Costa and Campilho *IPSJ Transactions on Computer Vision and Applications* (2017) 9:10
DOI 10.1186/s41074-017-0023-6

IPSJ Transactions on Computer Vision and Applications

EXPRESS PAPER Open Access

Convolutional bag of words for diabetic retinopathy detection from eye fundus images

Pedro Costa[1*] and Aurélio Campilho[1,2]

Abstract

This paper describes a methodology for diabetic retinopathy detection from eye fundus images using a generalization of the bag-of-visual-words (BoVW) method. We formulate the BoVW as two neural networks that can be trained jointly. Unlike the BoVW, our model is able to learn how to perform feature extraction, feature encoding, and classification guided by the classification error. The model achieves 0.97 area under the curve (AUC) on the DR2 dataset while the standard BoVW approach achieves 0.94 AUC. Also, it performs at the same level of the state-of-the-art on the Messidor dataset with 0.90 AUC.

Keywords: Bag-of-visual-words, Convolutional neural networks, Diabetic retinopathy detection

1 Introduction

Diabetic retinopathy (DR) is a complication of diabetes mellitus, wherein micro aneurysms start to form in the tiny vessels of the retina. In later stages of the disease, some retinal blood vessels may become blocked causing vision loss. Patients often do not have symptoms of the disease in its early stages which makes early diagnosis hard.

DR is the leading cause of blindness and visual loss in the working age population and the second most common cause in the USA [1]. Early detection of diabetic retinopathy is paramount for the success of the treatment, as it can prevent up to 98% of severe vision loss [2].

One way of performing the diagnosis of DR is by visually inspecting eye fundus images in order to detect retinal lesions. Examples of eye fundus images taken from the Messidor [3] dataset can be seen in Fig. 1. Although there are several grades of DR, we are only interested in the task of detecting the disease.

This work poses the task of discriminating between normal and pathological eye fundus images as a Multiple Instance Learning (MIL) problem. In the MIL task, each training example (called bag) is a set of feature vectors (called instances). Each bag has an associated label, but the labels of the instances are unknown.

The Standard Multiple Instance Learning assumption states that an example is positive if and only if one or more of its instances are positive [4]. Both normal and pathological eye fundus images contain several anatomical structures in common such as the macula, optical disk, and blood vessels. Nonetheless, only the pathological examples contain microaneurysms or any other lesion.

We pose the widely used bag-of-visual-words (BoVW) [5] method as a neural network, which allows it to refine the feature extraction and clustering functions by backpropagating the classification error.

We evaluated our method on the DR1 [6], DR2 [6], and Messidor [3] datasets. Our method obtained the new best results on the DR2 dataset and comparable results to the state-of-the-art on the Messidor dataset. To the knowledge of the authors, this is the first time that the DR1 dataset is used for the detection of DR.

Our contributions are as follows: a generalization of the BoVW method that outperforms the classical BoVW with a smaller number of visual words; we do not use lesion level information; and our method is more general than the classical approaches without compromising the results.

2 Related work

Most of the published work relies heavily on classical image processing methods and focuses on detecting individual DR lesions such as microaneurysms [7], drusen,

*Correspondence: pvcosta@inescporto.pt
[1] INESC TEC, Porto, Portugal
Full list of author information is available at the end of the article

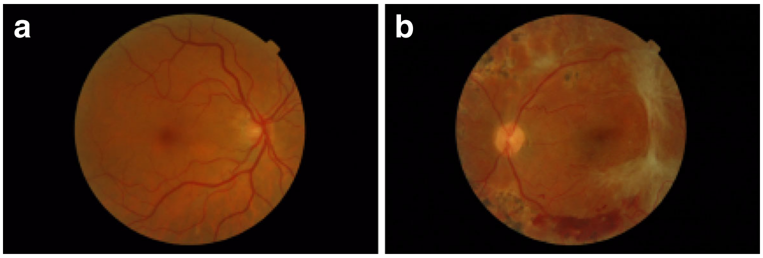

Fig. 1 Examples of eye fundus images of an healthy retina (**a**) and a retina with diabetic retinopathy (**b**). **a** Normal retina. **b** Pathological retina

exudates, and cotton-wool spots [8]. These methods typically follow a similar pipeline: image preprocessing, candidate extraction, and candidate classification. As each algorithm deals with a single lesion, a DR referral system has to combine the outputs of different methods to make a decision.

Amores [4] divided Multiple Instance Learning algorithms in three paradigms: Instance-Space (IS), Bag-Space (BS), and Embedded-Space (ES). The author compared these paradigms and found that the BS and ES paradigms have consistently better results than the IS one.

The IS paradigm assumes that each instance has discriminative power and the classifier is trained on the instance level. Then, for a new bag, the instance-level scores are aggregated to provide a final score. The BS paradigm assumes that the relevant information lies at the bag level. Since a bag is a non-vectorial entity, as it consists of a set of points, we need to define a distance function capable of comparing two sets of points. The ES paradigm maps each bag into a single feature vector that provides relevant statistics for the whole bag. The BoVW method falls into this category.

Pires et al. [6] applied the BoVW in the context of lesion classification in retinal images. The authors tried different feature extraction schemes and different coding and pooling functions. They found that in most cases, the best results were obtained by extracting and describing sparse features with Speeded Up Robust Features (SURF) [9], using semi-soft assignment as the coding function and the max function for the pooling operation.

Yan et al. [10] proposed a two-stage MIL method for computer tomography body part recognition. The authors divide the input image into patches and train a Convolutional Neural Network (CNN) on each patch on an IS fashion to find the discriminative patches. The second stage uses the learned discriminative patches as ground truth and adds a new class to the final layer to represent the non-discriminative patches. The image label corresponds to the label of the most discriminative patch.

On the other hand, Hou et al. [11] used a CNN on patches of gigapixel Whole Slide Tissue Images to differentiate between cancer subtypes. The authors start by dividing the image into patches and classify each patch into discriminative/non-discriminative using a CNN and expectation maximization. They then use the patch-level predictions to create an image level histogram that is used to train a logistic regression classifier. This is an ES method.

3 Methods

3.1 Bag-of-visual-words

The BoVW follows a specific pipeline: (i) extract local features from the images, (ii) learn a visual dictionary, (iii) create mid-level representations of the images using the visual dictionary, and (iv) learn a classifier using the mid-level representations. The visual dictionary consists of a set of M centers $c_m \in \mathcal{R}^D$ called visual words and is typically learned with K-means.

Following the terminology of Precioso and Cord [12], the extraction of local features from an image results in an unordered set of local descriptors named bag-of-features (BoF) $\mathcal{X} = x_i,\ i \in 1, \ldots, N$, where $x_i \in \mathcal{R}^D$ is a descriptor of a local interest point and N is the number of interest points detected in the image. As the number of interest points extracted varies from image to image (N will be a function of the image), the image does not have a fixed size feature vector, and therefore, it is not possible to directly apply a classifier.

A two-step pipeline is applied to each BoF in order to obtain its mid-level representation: coding and pooling. The coding step is a function $f : \mathcal{R}^D \rightarrow \mathcal{R}^M$ that maps

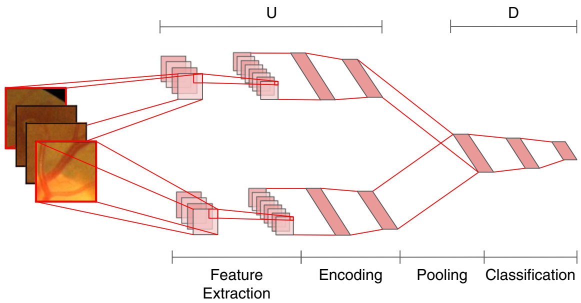

Fig. 2 A model that generalizes the BoVW. It is able to learn how to extract features, encode them, and classify the image

Table 1 Architecture of the network *U* when SURF is used on the DR1 and DR2 datasets

Layer	Output size
Input (dropout 30%)	128
FC-ReLU (dropout 50%)	500
FC-Softmax	100

a descriptor from the feature space into a representation based on the visual dictionary.

The pooling step aggregates the projections of the input features onto the dictionary to get a single representation. It can be represented by a function $g : \{w_i\}_{i \in 1,\dots,N} \rightarrow \mathcal{R}^M$ as $g(\{w_i\}) = z$. The max-pooling function performs well with the MIL assumption: the presence of a single microaneurysm is enough to classify the image as pathological:

$$g\left(\{w_i\}\right) = \max_{i \in \{1,\dots,N\}} w_{i,k},\ \forall k, \tag{1}$$

where w_i is the output of $f(x_i)$ and k is in the range $[\,0, M]$. The histogram z is then used as the feature vector of the image and used to train a classifier.

3.2 Convolutional bag-of-visual-words

The main problem with the BoVW approach is that the feature extraction, feature encoding, and classification are three separate problems. In order to counteract this problem, the size of the dictionary is increased to better divide the feature space, in some cases reaching hundreds of thousands of visual words. We created a neural network that is able to perform the same function as the BoVW but is able to learn jointly the feature extraction, coding, and classification functions.

Two networks are defined (Fig. 2): (i) a coding network $U(x; \theta_u)$ parameterized by θ_u that learns to extract features and cluster input instances together and (ii) a classification network $D(x; \theta_d)$ parameterized by θ_d that discriminates between normal and pathological mid-level representations.

The input of the model is convolved with U, resulting in a vector of latent variables, analogous to visual words. These latent variables, ideally, represent the different anatomical structures of the retina:

$$U(x_i) = p\left(w_i | x\right) \tag{2}$$

Table 2 Architecture of the network *U* when SURF is used on the Messidor dataset

Layer	Output size
Input (dropout 30%)	128
FC-ReLU (dropout 50%)	150
FC-Softmax	25

Table 3 Architecture of the network *U*

Layer	Filter size, stride	Output size
Input	-	64 × 64 × 1
Conv-ReLU	5×5, 1	60 × 60 × 16
Max-pool	2×2, 2	30 × 30 × 16
Conv-ReLU	3×3, 1	28 × 28 × 16
Max-pool	2×2, 2	14 × 14 × 16
Conv-ReLU	3×3, 1	12 × 12 × 16
Max-pool	2×2, 2	6 × 6 × 16
Conv-ReLU	3×3, 1	4 × 4 × 16
Max-pool	2×2, 2	2 × 2 × 16
Flatten	-	64
FC-Softmax	-	32

The classification network D receives as input a summary of the whole image and performs the classification. For instance, if the max-pooling function is used as the pooling function g, D decides based on which latent variables are present in the image and which are not. For DR detection, if one latent variable becomes active when a microaneurysm is present, D classifies the image as pathologic.

$$z = g\left(\{w_i\}\right) \tag{3}$$

$$D(z) = p(y|z) \tag{4}$$

The output of the model is, then, computed by $D\left(g(\{U(x_i)\})\right)\ \forall\, x_i \in \mathcal{X}$.

The function g is required to be differentiable (or almost everywhere differentiable), in order to train the two networks jointly. To train the network with back propagation, $\frac{\partial g(U(x_i))}{\partial U(x_i)}$ must be defined in order to update θ_u:

$$\frac{\partial \mathcal{L}}{\partial \theta_u} = \frac{\partial \mathcal{L}}{\partial z} \cdot \sum_i^N \left(\frac{\partial g\left(U(x_i)\right)}{\partial U(x_i)} \cdot \frac{\partial U(x_i)}{\partial \theta_u} \right), \tag{5}$$

where $\mathcal{L}$ is the loss function. Popular pooling functions like sum pooling and average pooling are differentiable and max pooling is almost everywhere differentiable and, as such, can be used with this model.

Similarly to BoVW, the model can receive bags of SURF as input, or any other numerical BoF $\mathcal{X}$. To do that, the

Table 4 The number of normal and pathological images in each dataset

Dataset	Normal	Pathological
DR1	595	482
DR2	337	98
Messidor	546	654

Table 5 Comparison of DR detection on the DR1 dataset

Method	AUC
Sparse SURF	93% ± 1%
Dense CNN	91%

feature extraction part of the network U is omitted and the BoF is given to the encoding part of U.

The advantage of this model over IS methods is that D is able to find relationships between the inputs. If the input instances are not discriminative, as when the output is the result of an XOR between two latent variables, this model is able to learn the classification function while IS methods cannot.

3.3 Procedure

The first step consists of extracting features from the image. We tried two strategies: (i) dense—extract patches from the image on a grid using different sizes and scales and (ii) sparse—extract SURF features from the image, since it has been empirically shown to have better results than other feature extraction methods on DR detection [6]. The OpenCV [13] implementation was used with default parameters.

After the extraction of local interest points, we proceed to describe them. Again, two strategies were used: (i) SURF—extracting the 128 dimensional extended feature vector and (ii) CNN—used only with dense features.

For the case when SURF was used on the DR1 and DR2 datasets, we used the network U as depicted in Table 1, and in Messidor, the used architecture is shown in Table 2. The network U in Table 3 was used for the CNN strategy. In both cases, D was a single fully connected layer. We used dropout [14], batch-norm [15], and dataset augmentation to regularize the network.

4 Evaluation

4.1 Datasets

We tested our model on three different datasets: (i) DR1[6]—grayscale 640 × 480 images. Images may be *Normal* or have one or more lesions. (ii) DR2[6]—grayscale 867 × 575 images. These images are divided by referral: images that indicate DR and normal images. (iii) Messidor [3]—RGB images labeled with the retinopathy grade, with 0 being normal and 1 to 3 being the different severity grades. Images have three different resolutions of 1440 × 960, 2240 × 1488, and 2304 × 1536.

Table 6 Comparison of DR detection on the DR2 dataset

Method	AUC
Pires et al. (2014) [6]	94%
Dense SURF	95% ± 1%
Sparse SURF	*97% ± 1%*
Dense CNN	*97%*

The italic entries were showing the best results

Table 7 Comparison of DR detection on the Messidor dataset

Method	AUC
Antal and Hajdu (2012) [19]	88%
Roychowdhury et al. (2014) [18]	*90%*
Quellec et al. (2015) [17]	89%
Sparse SURF	*90%*

The italic entries were showing the best results

We were only interested in distinguishing between normal and pathological images, so all images from DR1 with lesions and all images from Messidor with grade ≥1 were considered pathological. The number of normal and pathological images in each dataset is listed in Table 4.

4.2 Results

We followed the same evaluation procedure on the three datasets: we held-out 20% of each dataset for testing, while 65% was used to train and 15% for validation. The values for the hyper-parameters were found using random search [16], choosing the values that had the best area under the curve (AUC). The results are shown in Table 5 (DR1), Table 6 (DR2), and Table 7 (Messidor).

Our method was able to achieve 93% AUC in the DR1 dataset extracting SURF features. We were expecting the CNN to perform better, but it only achieved 91% AUC.

Pires et al. [6] used a BoVW with 1000 visual words and achieved 94% AUC on the DR2 dataset, while our method, with only 100 visual words, was able to achieve 97% (Table 6). Quellec et al. [17] also used a variation of the BoVW, with a more complex encoding scheme, achieving 89% AUC on the Messidor dataset, while ours achieved 90% (Table 7).

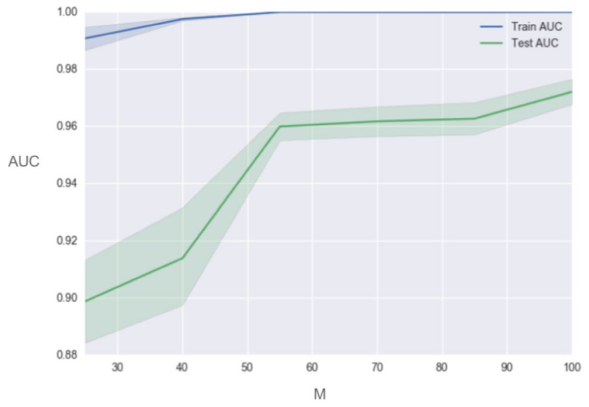

Fig. 3 Results of the model on the DR2 dataset by varying the number of visual words

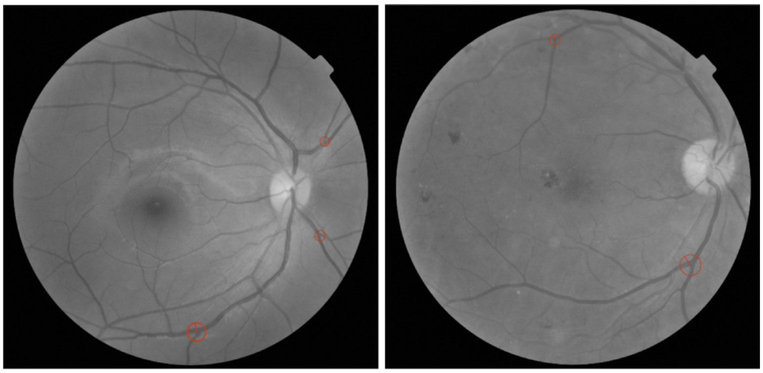

Fig. 4 Visual word that becomes active on vessel bifurcations. Appears both on normal (*left*) and pathological (*right*) images. Best viewed in *color*

Our method is also able to obtain comparable results to Roychowdhurry et al.'s [18] method, that relies on feature engineering.

We also tested the impact of the number of visual words on the results of the model by training the network using sparse SURF features on the DR2 dataset while varying M (Fig. 3). With 55 visual words, the model already achieves 96% AUC and then slowly increases to 97% AUC with 100 visual words. We did not see any improvements on the test set AUC by using more than 100 visual words.

Since we used 100 visual words to train the Sparse SURF model, it is easy to inspect what the model learnt. The different visual words are still divided by their visual appearance. We wanted to see if the model was indeed capable of learning the different anatomical structures of the retina.

By looking at the instances that become active at each visual word, it was possible to confirm that the model still divides the instances by their visual similarity. For instance, there is one visual word that becomes active on blood vessel intersections, as seen in Fig. 4, and another on the macula, although, there are other visual words that are not as interesting, such as one that becomes active on points on the border of the image.

Nonetheless, there are some visual words that are only active on pathological images. One of such visual words is shown in Fig. 5 and becomes active on bright lesions.

5 Conclusions and future work

We presented a neural network architecture that generalizes the well-known BoVW model. It is capable of using existing feature extraction methods or to extract features from images using a CNN. We do not encode any prior knowledge into the model, resulting in a very general method, without sacrificing the performance. Our method outperforms the BoVW and is comparable to the state-of-the-art approaches.

Since our method is able to learn with fewer number of visual words than the traditional BoVW approaches, it should be more interpretable. In the future, we want to evaluate the interpretability of the model.

We were expecting that using a CNN to extract features from the images would perform better than using SURF features, but in the case of the DR1 dataset, that

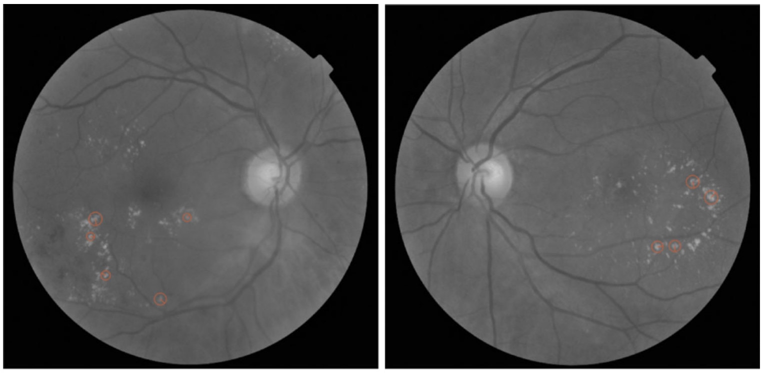

Fig. 5 Visual word that becomes active on bright lesions. Appears only on pathological images. Best viewed in *color*

was not true. These results might be due to the increased difficulty in optimizing the hyper-parameters with CNNs. It remains as future work to perform further tests with CNNs and evaluate the effects of the patch size on the results.

Acknowledgements
This work is financed by the ERDF, European Regional Development Fund, through the Operational Programme for Competitiveness and Internationalisation, COMPETE 2020 Programme, and by National Funds through the FCT, Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology), within project CMUP-ERI/TIC/0028/2014.

Authors' contributions
PC implemented the code, carried out the experiments and wrote the manuscript mainly. AC contributed to concept and wrote the manuscript partially. Both authors reviewed and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1] INESC TEC, Porto, Portugal. [2] Faculty of Engineering, University of Porto, Porto, Portugal.

Received: 21 February 2017 Accepted: 15 March 2017
Published online: 24 March 2017

References

1. Abramoff MD, Garvin MK, Sonka M (2010) Retinal imaging and image analysis. IEEE Rev Biomed Eng 3:169–208. doi:10.1109/RBME.2010.2084567
2. Economics, Access (2009) Future Sight Loss UK 1: The economic impact of partial sight and blindness in the UK adult population. RNIB
3. Decencière E, Zhang X, Cazuguel G, Lay B, Cochener B, Trone C, Gain P, Ordonez R, Massin P, Erginay A, Charton B, Klein JC (2014) Feedback on a publicly distributed image database: the Messidor database. Image Anal Stereology 33(3):231–234. doi:10.5566/ias.1155
4. Amores J (2013) Multiple instance classification: review, taxonomy and comparative study. Artif Intell 201:81–105. doi:10.1016/j.artint.2013.06.003
5. Sivic J, Zisserman A (2003) Video google: A text retrieval approach to object matching in videos. In: Iccv Vol. 2. pp 1470–1477
6. Pires R, Jelinek HF, Wainer J, Valle E, Rocha A (2014) Advancing bag-of-visual-words representations for lesion classification in retinal images. PLoS ONE 9(6):96814. doi:10.1371/journal.pone.0096814
7. Kamel M, Belkassim S, Mendonca AM, Campilho A (2001) A neural network approach for the automatic detection of microaneurysms in retinal angiograms. In: Neural Networks, 2001. Proceedings. IJCNN'01. International Joint Conference on. IEEE Vol. 4. pp 2695–2699. doi:10.1109/IJCNN.2001.938798
8. Niemeijer M, van Ginneken B, Russell SR, Suttorp-Schulten MSA, Abramoff MD (2007) Automated detection and differentiation of drusen, exudates, and cotton-wool spots in digital color fundus photographs for diabetic Retinopathy diagnosis. Invest Opthalmology Vis Sci 48(5):2260. doi:10.1167/iovs.06-0996
9. Bay H, Tuytelaars T, Van Gool L (2006) SURF: speeded up robust features. In: Computer Vision–ECCV 2006. pp 404–417. doi:10.1007/11744023_32
10. Yan Z, Zhan Y, Peng Z, Liao S, Shinagawa Y, Zhang S, Metaxas DN, Zhou XS (2016) Multi-instance deep learning: discover discriminative local anatomies for bodypart recognition. IEEE Trans Med Imaging 35(5):1332–1343. doi:10.1109/TMI.2016.2524985
11. Hou L, Samaras D, Kurc TM, Gao Y, Davis JE, Saltz JH (2015) Patch-based convolutional neural network for whole slide tissue image classification. arXiv preprint arXiv: … 7. 1504.07947
12. Precioso F, Cord M (2012) Machine learning approaches for visual information retrieval. In: Visual Indexing and Retrieval. Springer. pp 21–40. doi:10.1007/978-1-4614-3588-4_3
13. Bradski G (2000) The OpenCV Library. Dr Dobbs J Softw Tools 25:120–125. doi:10.1111/0023-8333.50.s1.10
14. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J Mach Learn Res 15:1929–1958. http://jmlr.org/papers/v15/srivastava14a.html
15. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv: 1502.03167
16. Bergstra J, Bengio Y (2012) Random search for hyper-parameter optimization. J Mach Learn Res 13(1):281–305
17. Quellec G, Lamard M, Erginay A, Chabouis A, Massin P, Cochener B, Cazuguel G (2016) Automatic detection of referral patients due to retinal pathologies through data mining. Med Image Anal 29:47–64. doi:10.1016/j.media.2015.12.006
18. Roychowdhury S, Koozekanani DD, Parhi KK (2014) DREAM: diabetic retinopathy analysis using machine learning. IEEE J Biomed Health Inform 18(5):1717–1728. doi:10.1109/JBHI.2013.2294635
19. Antal B, Hajdu A (2012) An ensemble-based system for microaneurysm detection and diabetic retinopathy grading. IEEE Trans Biomed Eng 59(6):1720–1726. doi:10.1109/TBME.2012.2193126. 1410.8577