Comparative Study of Convolutional Autoencoders and ORB Feature Descriptor for Street Image Similarity

Bogdan Bîndilă
m.bindila@student.utwente.nl
University of Twente
Enschede, Netherlands

Mark Bruderer
m.a.bruderervanblerk@student.utwente.nl
University of Twente
Enschede, Netherlands

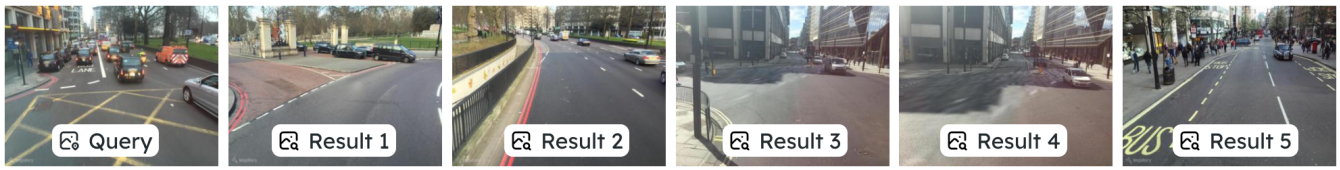

Figure 1. Output of our method

1 Introduction

Access to pertinent data is critical for the decision-making of both humans and automated agents. As the sizes of datasets continue to increase, so does the need for precise information retrieval (IR) systems. In the field of search engines and self-driving cars the requirement of searching for similar images in a dataset based on a query image arises.

The field of Image retrieval has two search methods, text-based image retrieval and content-based image retrieval (CBIR). The former relies on metadata or labels, while the latter is based on finding visually similar images.

In this work, we focus on CBIR systems, which have evolved tremendously over the years. In this setting, a user has an information need, which in this case is an image, for which similar images from the collection should be retrieved. The general framework is composed of an optional image preprocessing stage, a feature extraction module, and a similarity-matching method.

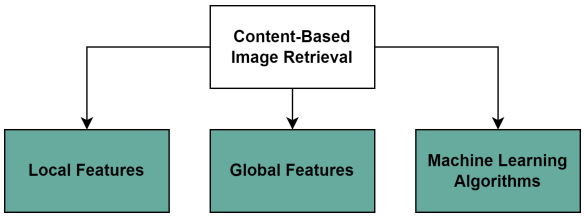

Figure 2. Feature extraction methods for CBID

The critical component in this type of system is the feature extraction module, which aims to convert human perception into a numerical representation [1]. Historically, many approaches have been studied, being grouped into three main categories as shown in Figure 2.

While local feature extractors have been used by many researchers because are invariant to scale and rotation, global features such as color, texture, shape, and spatial information need to be combined to offer a satisfactory performance.

CBIR methods often suffer from the *semantic gap* which means that low-level features do not give the same information as high-level concepts understood by humans. This leads to inaccurate image retrieval. Machine learning and particularly neural networks provide a way to automatically extract meaningful features from images, without human intervention. In this category fall supervised and unsupervised approaches. Lately, deep learning techniques have been used to bridge the semantic gap [1].

Considering that classical local feature descriptor methods like ORB are scale-invariant and do not close the semantic gap [2], we propose an improved approach to feature extraction based on convolutional autoencoders (CAE). This deep learning model can be trained in an unsupervised fashion to generate meaningful and compact representations of the input data [3].

We state the following null and alternative hypotheses:

Hypothesis H_0: Image encodings learned by means of a convolutional autoencoder does not improve the precision of image retrieval compared to image features extracted with ORB and K-means clustering.

Hypothesis H_1: Image encodings learned by means of a convolutional autoencoder improves the precision of image retrieval compared to the image features extracted with ORB and K-means clustering.

2 Related Work

The problem of extracting meaning from images has been studied by many authors. Alex Krizhevsky and Geoffrey Hinton [3] trained an autoencoder to reduce images to binary, 256 bit codes. The hamming distance is then used to find relevant images for a "hashed" query image. This paper puts forward the advantage of fast bit-wise operations and

Figure 3. Example of original and resized image

efficient storage of these binary encodings. The binary encodings can also be used as "semantic" memory addresses, this makes the retrieval time complexity constant instead of linearly proportional to the size of the dataset.

To classify the images of galaxies from telescopes Seo et al. [4], also used a convolutional autoencoder. The similarity between images of galaxies were found by using the Euclidean distance between encoded images. The model in this study was based on the ResNet-18 [5] architecture. The autoencoder was created by stacking a ResNet-18 network followed by a reversed ResNet-18 network. By using this method the training of this classification system could be done in a self-supervised way, only requiring 1 human annotated label for each reference galaxy image.

Autoencoders (AE) represent the solution chosen for feature extraction in the context of medical image analysis and search due to their performance and unsupervised training method. In [6], AEs are trained to encode x-ray images to binary codes as a replacement for local feature descriptors that do not detect well the corners in this use case. A series of stacked AEs are trained in a greedy fashion to meet the needs of this problem.

Moreover, convolutional autoencoders are effective feature extractors from pulmonary computed tomography (CT) images [7]. Chen et al. employ a deep-learning-based technique to avoid the uncertainty of hand-crafted features for similar lung nodule retrieval. After having the features, the similarity between nodules is computed by training a separate network with pairs of similar and dissimilar images.

3 Methods

3.1 Data Description

We use a subset of the Mapillary Street-level Sequences Dataset [8] to compare a local feature descriptor as ORB [2] with a modern global feature extractor based on deep learning techniques. It represents the vastest dataset for place recognition, containing 1.6 million images from 30 cities taken in a multitude of environmental conditions. In our case, we utilize 1,500 images taken on the streets of London. We train the model without using the local metadata. The dataset is divided into 500 query images and 1,000 map images. For each query instance, there are between 1 and 30 relevant map pictures, the relevance judgments being stored in a matrix of similarities. Our focus is on retrieving relevant results without using this degree of similarity. The pictures have three color channels, a height of 256 pixels, and a variable width of 341, 367, or 455 pixels.

The data pre-processing consists of two simple steps: a downscaling to a standardized size of 256 by 336, followed by a pixel-level scaling between 0 and 1. This specific downscaling is necessary because each dimension has to be a multiple of 16 to ensure that the output image from our convolutional autoencoder has identical dimensions as the input image. Figure 3 illustrates the resizing of a larger image from our dataset. While the aspect ratio is not maintained, the scene details remain intact.

3.2 Model Selection

We propose an end-to-end unsupervised approach for extracting features in an automated way based on CAE. This type of feed-forward network is trained to produce an approximation of the identity mapping between inputs and outputs using backpropagation [9]. The central idea of an autoassociative neural network is represented by the bottleneck between input and output that force the learning of a compressed representation of the input. This compact representation captures the essential features of the input data by suppressing noise and irrelevant details. It can be used to compute similarities between images because they are much more semantically meaningful [3], narrowing the semantic gap between input images and their representations.

Compared to the fully-connected autoencoder, convolutional autoencoders are more suitable for CBIR applications due to their invariance to translation, scale, and rotation [1]. These properties make them a powerful tool for addressing our image retrieval challenge.

3.2.1 Architectural Choices.

To compare our approach with the ORB local feature extractor, we tried to match as closely as possible the number of 12,800 values that are composed by the 50 ORB extracted features of 256 values. Consequently, the architecture is composed of 4 convolutional layers in the encoder and 4 transposed convolutional layers plus a fifth convolutional layer in the decoder. All operations except the last one use zero padding, a stride of 2 and, $ReLU(x) = max(0, x)$ activation function. The last convolutional layer has the sigmoid $\sigma(x) = \frac{1}{1+e^{-x}}$ activation function to map the output pixel values between 0 and 1 and 3 filters to map the data back to 3 channels.

Figure 4 illustrates the architecture and details the output shapes and number of filters used in each layer. When it comes to the bottleneck, the model reduces the input to 12,768 floating point values. This *compression ratio of 20.21 times* is extremely high compared to the results presented in [6] where the compression was only 74.61%. In terms of trainable parameters, it has only 47027.

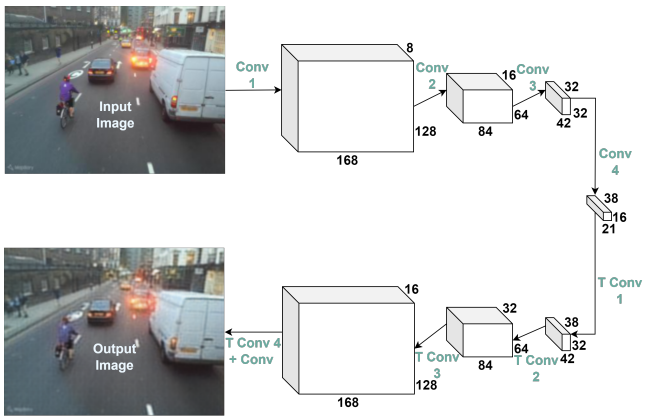

Figure 4. Architecture of the Convolutional Autoencoder

We experimented by enhancing the network with max pooling and batch normalization layers as shown in [7], but we did not notice any improvement in the loss function, so we removed them to keep the network as light as possible.

The autoencoder is trained through backpropagation to minimize the mean squared error $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ between the input and reconstructed image as in [3]. The process is optimized by using the Adam optimizer with a fixed learning rate.

3.2.2 Hyperparameter Optimization.

Now having defined the main components of our model, we further want to optimize the batch size and learning rate. We performed a grid search over the following space:

- **batch size:** 1, 16, 32, 64, 128
- **learning rate:** 0.0001, 0.0005, 0.001, 0.005, 0.01

Each model instance is trained on 90% of the available data for at most 100 epochs and validated on 10% of the remaining images. However, the procedure could end earlier due to the early stopping if the validation loss does not increase for 3 consecutive epochs.

After the optimization step, the convolutional autoencoder is retrained on the entire dataset of 1,000 map images with the best combination of batch size, learning rate, and number of epochs.

3.3 Evaluation

After the model is trained we can evaluate the performance of the proposed content-based image retrieval system. In order to retrieve results based on a query image, all images must first be encoded. Then our system returns the images in the database, which have the smallest distance to the query image. The distances are calculated using only the encoded vectors.

We first obtain encodings by doing a forward pass of the encoder for each image and flattening the output of the model into a vector. These encodings are saved into two files for later use. Each *encoding has the size of* 12768. The map file contains 1000 vectors and the query file contains 500 vectors.

To retrieve images from the database for a given query we first compute the *Euclidean distance* between the encoded query image and each encoded image in the map dataset. After sorting the Euclidean distances we can rank the similarity of the map images to the query image.

To evaluate the performance of the content-based retrieval, we recorded the *Average Precision* (equation 1) and *R-Precision* (equation 2) metrics.

$$AP_q = \frac{1}{|R|}\sum_{i}^{|Retrieved|} relevant(i) \cdot Precision_i \tag{1}$$

$$rP_q = \frac{1}{|R|}\sum_{i}^{|R|} 1 \cdot relevant(i) \tag{2}$$

We calculated $AP_q\ for\ q \in 1 \ldots 500$ with method A (convolutional autoencoder) and $AP_q\ for\ q \in 1 \ldots 500$ with method B (ORB feature descriptor). This yields the samples AP_{qA} and AP_{qB} of average precision. We also compute the paired differences of these two samples $Z_q = AP_{qA} - AP_{qB}$.

In order to perform a test of our Null Hypothesis, we first went through the task of choosing an appropriate statistical test. We considered as possible statistical tests, the *Student's t-test* [10], *Wilcoxon signed-rank test* [11] and the *Sign test* [12]. These tests were considered because they are mentioned as suitable tests in the context of IR systems [13].

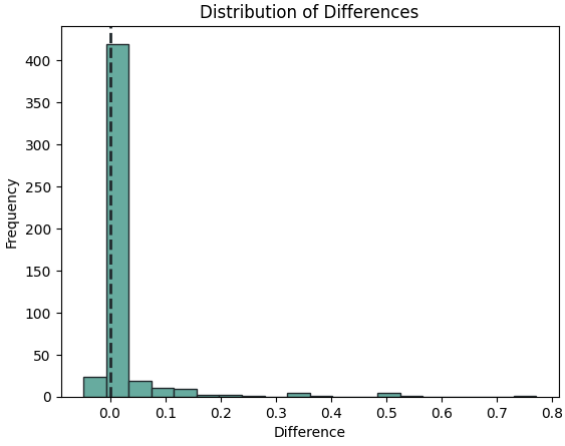

Figure 5. Histogram of Paired Precision Differences

The differences in paired observations did not seem to follow a normal distribution visually (See figure 5). We performed a *Shapiro-Wilk* test and found that the null hypothesis of the test should be rejected which made us conclude that the assumption of normality could not be justified. This eliminates the possibility of using a Student's t-test as it assumes that the distribution is Gaussian.

We then checked the assumption for the *Wilcoxon signed-rank test*. This test assumes that the differences of paired observations should be symmetrical to the median, we checked this visually, from which we saw that the observations are skewed to the right of the median.

For the Sign Test, we only assume that the differences are independent and come from the same population, i.e. the population of differences in average precision. Additionally, both AP_{qA} and AP_{qB} have the same ordinal scale (average precision scale).

We chose a significance level of $\alpha = 0.05$. Since our null hypothesis is that the image encodings do not improve the performance, we use the right-tailed variant of the sign test. The outcome of this test is presented in the results section.

4 Results

We evaluate various instances of the CAE based on the validation loss. The best-performing instance is obtained training for *50 epochs* with a *learning rate of 0.0005* and using the stochastic gradient descent.

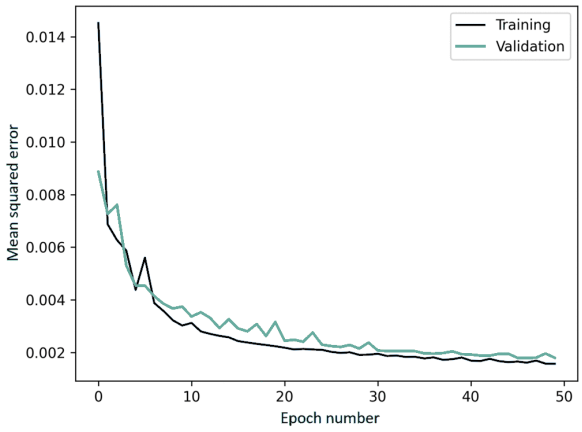

Figure 6. Losses evolution of the optimized model

Figure 6 depicts the training and validation loss obtained after training the model with the aforementioned hyperparameters. Even though a *batch size of 1* is used, the loss functions are very smooth. The loss values after 50 epochs are *0.001794 for validation* and *0.001568 for training* sets. When the model was re-trained on the entire map dataset, the *loss function reached 0.0014*, proving that the model learned even more from the newly added examples.

Figure 4 contains an actual reconstructed image and it can be observed that from a qualitative perspective, all shapes and colors are recreated almost perfectly. The information loss is at a texture level that is altered throughout the entire image. This example is representative for the entire set, the rest of them having a similar reconstruction quality.

4.1 Statistical Test

We performed the *Sign-Test* for the two methods using the 500 paired differences of precision. We obtained the p-value of $2.952 \cdot 10^{-14}$. Based on our chosen significance value of 0.05 we reject the null hypothesis of the right-tailed Sign Test. This means that the precision in method A tends to be higher than the precision of method B. The average values and variances of these metrics are presented in Table 1.

	$\bar{x_A}$	s_A^2	$\bar{x_B}$	s_B^2
Average Precision	0.029	0.006	0.007	0.00004
R-Precision	0.046	0.020	0.010	0.002

Table 1. Mean and variance for Average Precision and R-Precision. Method A (ours) and Method B (ORB)

5 Discussion

The convolutional autoencoder method to represent images provides a good way to bridge the semantic gap. By creating a bottleneck the model must learn to compress the image into a higher level features. This representation can be used to retrieve semantically close images using a distance metric.

As can be seen in the output of our system (Figure 7), we retrieved for the given query image the images that come from the same road. We can also see in this figure that the first image is only 6.09 meters away from the query image. As the rank of the image increases the physical distance increases. The fifth image is not a relevant image in this example. The system performs well here probably due to the closeness in color between the images. In the fifth image, there is a similar level of brightness which might cause this false positive.

The best-performing query for our method is shown in Figure 8. In this example, the precision at 5 is a perfect 1.0. This again is in the scene with a dominant brown color in the image. We compared this to the same query by using the ORB feature extractor (figure 9). The ORB method does not yield any relevant images.

ORB performs the best in a query image coming from a park (figure 10). It retrieves two relevant images in the top 5 coming from under 20 meters away. By contrast, our method does not retrieve the relevant images. If we examine these two examples visually we can see that the ORB method uses the trees to correctly retrieve the two relevant images. Our method also returns images with similar-looking trees but they do not match the original location.

We rejected our null hypothesis by means of a sign test. A drawback of the sign test is that is not considered statistically powerful compared to other tests. Another important disadvantage is that only the direction of the difference and not the magnitude are considered.

6 Conclusion

We trained a convolutional autoencoder with the goal of extracting features from images. This model was trained in an unsupervised way. We used the trained encoder to create encodings for the images in our dataset. This allowed us to compute distances between the images in the dataset. We then evaluated the performance of this approach and compared it to the approach using ORB feature extractors and K-Means clustering.

We reject our stated null hypothesis that image encodings learned by means of a convolutional autoencoder do not improve the precision of image retrieval compared to image features extracted with ORB and K-means clustering. This supports the reasoning that machine learning-based feature extractors such as a convolutional encoder produce better results due to the inclusion of general aspects of the image such as color, texture, and spatial connections.

During this study, we have not fully experimented with various architectures. Thus, a natural step would be to quantify the implications of training more complex networks and compute the correlation between various compression rates and the performance of the retrieval system. Moreover, to improve the performance we consider that data augmentation techniques can help the model generalize better. Last but not least, training with grayscale images can speed up the learning process and reduce the computational cost without a significant performance decrease.

References

[1] Ibtihaal M Hameed, Sadiq H Abdulhussain, and Basheera M Mahmmod. "Content-based image retrieval: A review of recent trends". In: *Cogent Engineering* 8.1 (2021), p. 1927469.

[2] Ethan Rublee et al. "ORB: An efficient alternative to SIFT or SURF". In: *2011 International conference on computer vision*. Ieee. 2011, pp. 2564–2571.

[3] Alex Krizhevsky and Geoffrey E Hinton. "Using very deep autoencoders for content-based image retrieval." In: *ESANN*. Vol. 1. Citeseer. 2011, p. 2.

[4] Eunsuk Seo et al. "Similar Image Retrieval using Autoencoder. I. Automatic Morphology Classification of Galaxies". In: *Publications of the Astronomical Society of the Pacific* 135.1050 (2023), p. 084101. DOI: 10.1088/1538-3873/ace851. URL: https://dx.doi.org/10.1088/1538-3873/ace851.

[5] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].

[6] S Sharma et al. "Stacked autoencoders for medical image search". In: *Advances in Visual Computing: 12th International Symposium, ISVC 2016, Las Vegas, NV, USA, December 12-14, 2016, Proceedings, Part I 12*. Springer. 2016, pp. 45–54.

[7] Min Chen et al. "Deep feature learning for medical image analysis with convolutional autoencoder neural network". In: *IEEE Transactions on Big Data* 7.4 (2017), pp. 750–758.

[8] Frederik Warburg et al. "Mapillary street-level sequences: A dataset for lifelong place recognition". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 2626–2635.

[9] Mark A Kramer. "Autoassociative neural networks". In: *Computers & chemical engineering* 16.4 (1992), pp. 313–328.

[10] Student. "The probable error of a mean". In: *Biometrika* 6.1 (Mar. 1908), p. 1. DOI: 10.2307/2331554. URL: https://doi.org/10.2307/2331554.

[11] Frank Wilcoxon. "Individual Comparisons by Ranking Methods". In: *Biometrics Bulletin* 1.6 (1945), pp. 80–83. ISSN: 00994987. URL: http://www.jstor.org/stable/3001968 (visited on 11/10/2023).

[12] John Arbuthnot. "II. An argument for divine providence, taken from the constant regularity observ'd in the births of both sexes. By Dr. John Arbuthnott, Physitian in Ordinary to Her Majesty, and Fellow of the College of Physitians and the Royal Society". In: *Philosophical Transactions of the Royal Society of London* 27.328 (1710), pp. 186–190. DOI: 10.1098/rstl.1710.0011. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rstl.1710.0011. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rstl.1710.0011.

[13] Mark D. Smucker, James Allan, and Ben Carterette. "A Comparison of Statistical Significance Tests for Information Retrieval Evaluation". In: *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*. CIKM '07. Lisbon, Portugal: Association for Computing Machinery, 2007, 623–632. ISBN: 9781595938039. DOI: 10.1145/1321440.1321528. URL: https://doi.org/10.1145/1321440.1321528.

A Example Queries

See next page.

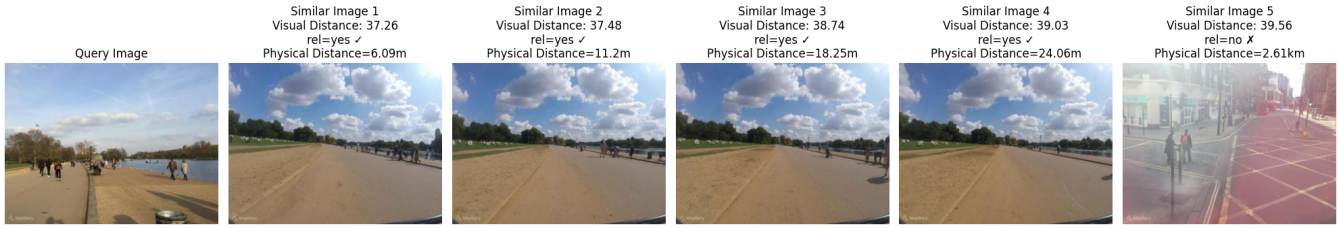

Figure 7. Result for (q=121)

Figure 8. Best Average Precision Run Of Our Method (q=122)

Figure 9. Orb Performance (q=122)

Figure 10. Best Average Precision ORB (q=223)

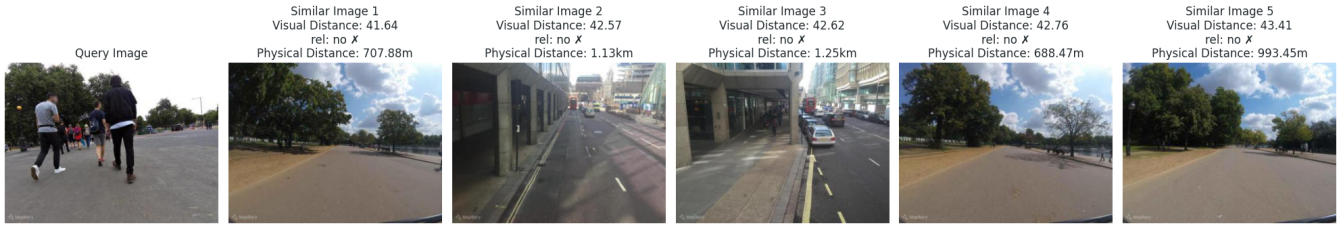

Figure 11. Our Method Performance (q=223)