Multimodal Object Recognition Using Deep Learning Representations Extracted from Images and Smartphone Sensors

Javier Ortega Bastida, Antonio-Javier Gallego(✉), and Antonio Pertusa

Department of Software and Computing Systems, University of Alicante,
03690 San Vicente del Raspeig, Alicante, Spain
job5@alu.ua.es,{jgallego,pertusa}@dlsi.ua.es

Abstract. In this work, we present a multimodal approach to perform object recognition from photographs taken using smartphones. The proposed method extracts neural codes from the input image using a Convolutional Neural Network (CNN), and combines them with a series of metadata gathered from the smartphone sensors when the picture was taken. These metadata complement the visual contents and they can provide additional information in order to determine the target class. We add feature selection and metadata pre-processing, by encoding textual features, such as the kind of place where a picture was taken, using Doc2Vec in order to maintain the semantics. The deep representations extracted from images and metadata are combined with early fusion to classify samples using different machine learning methods (k-Nearest Neighbors, Random Forests and Support Vector Machines). Results show that metadata preprocessing is beneficial, SVM outperforms kNN when using neural codes on the visual information, and the combination of neural codes and metadata only improves the results slightly when the images are classified into very general categories.

Keywords: Multimodality · Object recognition · Metadata · Learning representations

1 Introduction

Object recognition is a field of computer vision that aims to identify objects or entities in images or videos. This is a highly active topic which can be particularly useful for mobile devices [7] as regards retrieving information about objects on the fly. Using supervised learning techniques such as Convolutional Neural Networks (CNN), we can build models to recognize the objects present in an image.

In order to achieve a better prediction, some recognition methods use additional information to help identify the predominant objects in images. In some cases, metadata such as the GPS location [15] are included. This leads to multimodal methods which use different information sources. Some previous

R. Vera-Rodriguez et al. (Eds.): CIARP 2018, LNCS 11401, pp. 521–529, 2019.
https://doi.org/10.1007/978-3-030-13469-3_61

approaches successfully combined visual descriptors with textual information [3], and also with features such as the camera metadata [2] in order to facilitate object identification. Multimodality in deep learning has also been studied for the creation of complex networks which can detect the most relevant characteristics of the different data sources. An example is the Multimodal Convolutional Neural Network [10] for matching images and sentences, or the Image-Text Multimodal Representation Learning by Adversarial Backpropagation [13].

In this work, we use the MirBot [15] dataset which contains images taken from smartphones along with their associated metadata. MirBot[1] is a collaborative object recognition system which allows users to take a photograph and select a rectangular region of interest (ROI) in which a target object is located. The image, the ROI coordinates and a series of associated metadata are sent to a server, which performs a similarity search and returns the class (a WordNet [4] synset such as chair, dog, laptop, etc.) of the most likely image in the training set. The app users can validate the system response in order to improve the classification results for future queries, and this feedback allows the database to grow continuously with new labeled images.

The metadata of the Mirbot dataset are extracted from the smartphone sensors (angle with regard to the horizontal, gyroscope, flash, GPS, etc.), reverse geocoding information (type of place, country, closest points of interest, etc.) and EXIF camera data (aperture, brightness, ISO, etc.). The gathered metadata can be used to reduce the search space. For instance, if a user takes a photograph of an elephant, it is more likely that it will be in a zoo rather than on a beach, that the angle respect to the horizontal will be close to 90°, and that the flash will be off [15].

In the present work, we extend the multimodal method from [15], and use a supervised learning classifier to perform early-fusion on the learned deep representations of both images and metadata.

The remainder of this paper is organized as follows. Section 2 describes the dataset and Sect. 3 the methodology used for multimodal classification. The evaluation results are detailed in Sect. 4. Finally, Sect. 5 addresses our conclusions and future work.

2 Dataset

As the MirBot data is dynamic and user-driven, statistics change over time. In the following, experiments refer to the dataset from October 23, 2016 for a direct comparison with the results given in [15]. On this date, $3,431$ users had added $25,292$ images distributed in $1,808$ classes. Some objects appear more frequently than others and the classes are, therefore, highly unbalanced. Most images are categorized as objects ($18,685$), followed by animals ($4,928$), food/drinks ($1,113$), and plants (546).

[1] http://www.mirbot.com.

2.1 Metadata

Device Metadata. 29 metadata are obtained from the smartphone sensors for each image as described in [14]. These metadata correspond to the device information (model, version, etc.), geolocation data (latitude, longitude, altitude, locality, sublocality, PC, country, etc.), activation of the camera flash, and the sensor values (accelerometer, gyroscope, network status, etc.).

Gisgraphy Features. In addition, given a latitude and a longitude, reverse geocoding is performed in the server with Gisgraphy[2], which uses the GeoNames geographical database. This allows to obtain valuable data such as the feature class and code [1] that provide information about the kind of place (for example, University, Park, Restaurant, Zoo, etc). The list of the 17 Gisgraphy features can be seen in [14].

EXIF Metadata. The camera parameters of the pictures are also stored. The exchangeable image file format (EXIF) information sent to the server includes 23 parameters such as the focal length, aperture value, brightness, ISO speed, white balance, etc.

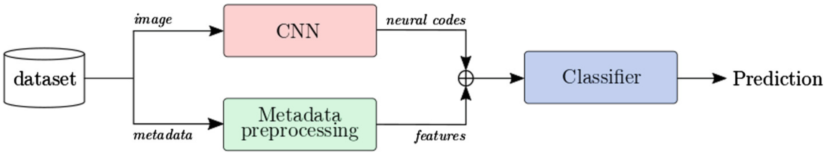


Fig. 1. Overall architecture of the proposed method.

3 Methodology

The proposed architecture for classification is summarized in Fig. 1. On the one hand, we send the input image to a CNN to generate the neural codes that represent the visual information. On the other hand, we use a series of metadata, which can either be numerical values (such as pitch, sharpness, etc.) or textual (such as country, gis feature code, gis feature class, etc.). Then, we concatenate the neural codes to the metadata features to be used as input for classification.

3.1 Neural Codes Extraction

Color images are resized to 224 × 224 pixels and given to a ResNet50 [6] CNN pre-trained with ImageNet and fine-tuned with the MirBot dataset. The visual features correspond to the neural codes (vectors of dimension 1,256) extracted from the last hidden layer of the CNN and normalized using ℓ_2. The details to get these visual descriptors are given in [15].

[2] http://www.gisgraphy.com/.

3.2 Metadata Preprocessing

MirBot metadata include numerical values, categorical data and text strings which have to be presented as sequential values to a classifier. In this work, like in [15], the features *osversion* and *model* are first removed, along with all the information related to an specific user such as its identifier.

Those metadata containing numerical values (such as pitch, sharpness, focal length, etc.) are normalized into the interval $[0, 1]$. In [15], textual metadata (such as country, gis feature code, gis feature class, etc.) were codified in a one-hot manner as they have not any specific ordering. This way, the distance between two categorical features can only be 1 if they are different or 0 if they match.

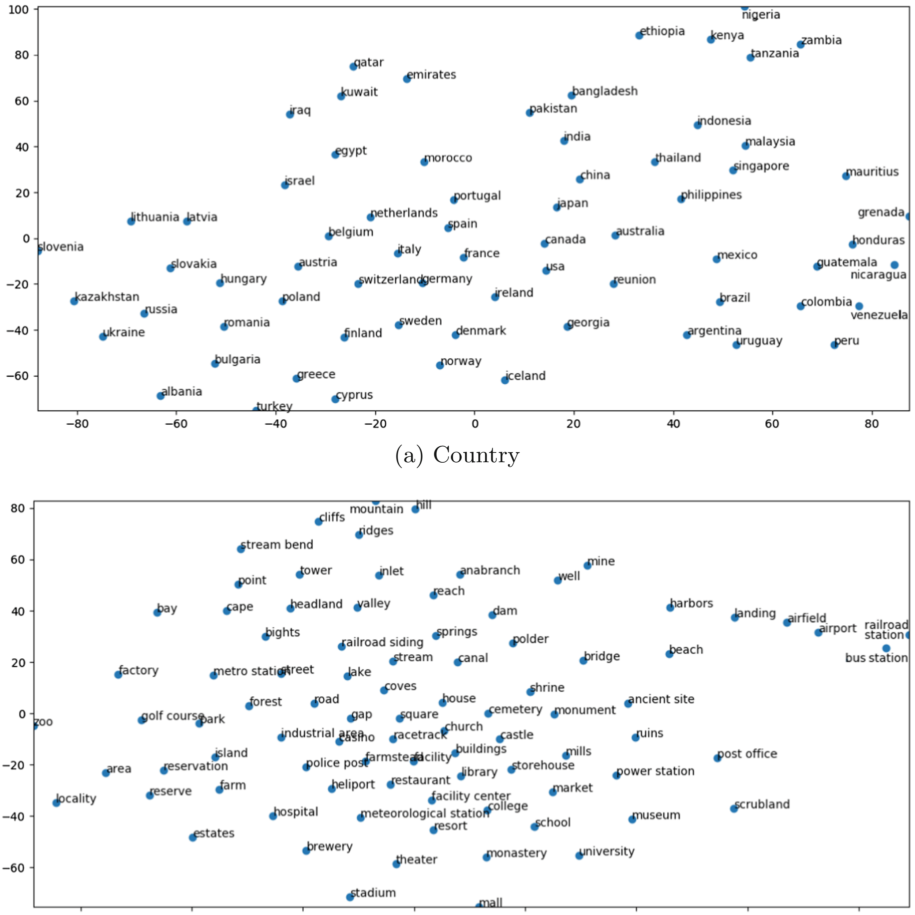
In this work, we pre-process the textual features. For example, there are some strings (such as the address) in many different languages. All these strings were translated into English. To automate this process we used the Google Translator API.

In addition, the problem of using a one-hot vector for representing textual data is that semantics are lost. In order to address this issue, differently from [15], in this work we propose to encode the categorical values using Doc2Vec [9], which is an extension of Word2Vec [12]. As its name suggests, Doc2Vec extracts a vector that represents the paragraphs and sentences, considering the context of the words in the paragraph. Doc2Vec is used instead Word2Vec because the textual strings are composed by sentences.

We used a Doc2Vec model [8] implemented in Gensim, a Python library for vector space modeling. This library includes two pre-trained models: *English Wikipedia DBOW* and *Associated Press New DBOW*. Initially, we tested the two models on our metadata with the default parameters, and the best accuracy was obtained using the Wikipedia model, consistently with the results given in [8].

Some of the metadata returned by the mobile device are codes instead of words (such as UK for United Kingdom in the country data or SCH for School in the feature code). In these cases, we determined that it was more effective to use the full name represented by the codes. In order to encode the feature codes (type of place where the picture was taken), we created a sentence by concatenating the name with its corresponding description given in [1]. For example, the code "SCH" is converted into "*School, building where instruction in one or more branches of knowledge takes place*" to be used as input for the Doc2Vec model.

With this pre-trained model, we transformed the following text features: name, locality, sublocality, admin-area, thoroughfare, gis-name, gis-adm1-name, gis-adm2-name, gis-adm3-name, country, gis-feature-code and gis-feature-class into vector embeddings. Figure 2 shows an example of the Doc2Vec processing with our dataset using t-Distributed Stochastic Neighbor Embedding (t-SNE [11]). It can be seen that similar concepts are grouped together. For example, country values of South America, Africa or Europe are close.



(a) Country

(b) Feature Code

Fig. 2. Document embeddings projection into a reduced space for country and feature codes using t-SNE [11].

3.3 Classification

Once the visual and metadata features are extracted, different classifiers can be used for this task. In particular, we evaluate k-Nearest Neighbors (kNN), Support Vector Machines (SVM) and Random Forests (RF). In the case of multimodal experiments, metadata features are appended to the neural codes to serve as input for the classifiers.

Different parameters for the classifiers were evaluated: kNN with $k \in [1, 100]$; SVM with $C \in [1, 1000]$; and RF with the number of trees within the range $[5, 1000]$.

4 Experiments

In this section, we evaluate the accuracy improvements offered by the new approaches presented in this paper with respect to the previous version of Mirbot [15]. We compare the results using metadata with one-hot encoding and with the preprocessed Doc2Vec model, the results with the visual features, and the combination of visual and metadata features.

Experiments were performed using a 5-fold cross validation. Only the images belonging to the classes with more than one prototype were used for evaluation ($24,794$ images from $1,180$ classes). The accuracy is provided using the top-1 evaluation metric, where a true positive is considered when the class of the closest prototype matches the query class. The classification was done at three levels: Root level (with the 5 main categories: animals, food and drink, man-made objects, natural objects, and plants), the second level of the WordNet hierarchy (with 92 classes), and the leaf level (with the 1,180 classes).

Evaluation Using Metadata. Attribute selection was first performed in order to rank and select the best subset of metadata features. For this, we applied several selection methods [5]: Best First, Genetic search, Greedy Stepwise, Linear Forward Selection, Random Search, Scatter Search V1, Subset Size Forward Selection, and InfoGain. After testing all these selection techniques, we applied a voting scheme to select the best attributes, which are shown in Table 1. The rest of the attributes were ignored for the following stages.

Table 1. Selected metadata using different attribute selection methods with a voting scheme. All features are numerical values except by those pre-processed using Doc2Vec, which are marked with (*).

Sensors	Location	EXIF
pitch	reliable location	sharpness
selected area	country (*)	focal length
wifi	ocean	brightness value
flash	gis feature code (*)	color space
	gis feature class (*)	subject area

As expected, one of the most representative metadata is the feature code [1], which stores the kind of place: Zoo, Mall, University, Beach, etc.

Table 2 shows the best results for each classifier. The best results with kNN were obtained with a very low neighbor value ($k = 1$). When using RF, the highest accuracy was obtained with 150–300 trees, and SVM did not improved the accuracy with values of C larger than 10. The results obtained for the first levels of the hierarchy are surprisingly good considering that the classification is performed without any visual information and there are $1,180$ classes. An explanation for this is that the dataset is highly unbalanced. We checked the confusion matrices in order to assess that the yielded classes are varied and there is no overfitting.

Table 2. Comparison of the best results for each classifier using the metadata without preprocessing and with preprocessing.

Method	Without preprocessing			With preprocessing		
	Root	2nd level	Class	Root	2nd level	Class
kNN	73.67	51.73	7.31	75.52	52.08	**27.37**
RF	67.80	35.31	9.94	**76.96**	52.77	20.01
SVM	73.70	52.01	6.29	76.09	**54.88**	15.51

Evaluation Using Visual Features. Results using Neural Codes (NC) are shown in Table 3. The kNN classifier was already evaluated in [15], but in the present work we include RF and SVM accuracy. As can be seen, RF outperforms the results from kNN given in [15] at the class level, although SVM obtains the best results at the root level.

Table 3. Comparison of the best results for each classifier using only the NC and the multimodal data (the combination of NC and metadata).

Method	Neural codes			Multimodal data		
	Root	2nd level	Class	Root	2nd level	Class
kNN	93.72	83.60	77.70	94.36	82.19	58.28
RF	90.24	**84.24**	**78.68**	90.60	80.11	51.94
SVM	94.81	84.02	76.93	**94.98**	82.02	52.33

Evaluation Combining Metadata and Visual Features. Although the main source of information is given by the image features, metadata could complement this information. In [15], metadata were only used when the confidence was low, that is when the difference of the distances between the first and second class returned by the visual classifier was small. Here, we perform early fusion for comparison, and the results show that multimodal data is more adequate than using only the visual features at the root level (particularly using SVM), although in the other levels results clearly decrease.

5 Conclusions and Future Work

In this work, we use visual features and metadata for object recognition. Feature selection was performed to get the most suitable metadata features, and we show that encoding textual features using Doc2Vec outperforms a one-hot representation, as similar locations are also close in the vector space. In addition, we combine metadata with visual features in a early-fusion approach, although results only outperformed visual features at the root level.

Results obtained preprocessing metadata with Doc2Vec show a considerable accuracy improvement compared to the one-hot encoding used in [15]. We also show that SVM outperforms the results obtained in [15] with kNN.

It should be noted that the combination of metadata with the neural codes slightly increases the results at the root level but significantly decreases with finer levels. This may be because the metadata contains very general information that only helps identifying the highest hierarchy level.

As future work, we plan to evaluate multimodal neural networks for learning more complex relationships between data in order to improve classification at the class level.

Acknowledgment. This work was supported by the Pattern Recognition and Artificial Intelligence Group (PRAIg) from the University of Alicante, Spain.

References

1. Geonames feature codes. http://www.geonames.org/export/codes.html. Accessed 14 June 2018
2. Boutell, M., Luo, J.: Beyond pixels: exploiting camera metadata for photo classification. Pattern Recognit. **38**(6), 935–946 (2005)
3. Dinakaran, B., Annapurna, J., Kumar, C.A.: Interactive image retrieval using text and image content. Cybern. Inf. Technol. **10**(3), 20–30 (2010)
4. Fellbaum, C.: WordNet: an electronic lexical database (1998). https://doi.org/10.1139/h11-025
5. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explor. Newsl. **11**(1), 10–18 (2009)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385 (2015)
7. Howard, A.G., et al.: MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. CoRR (2017)
8. Lau, J.H., Baldwin, T.: An empirical evaluation of doc2vec with practical insights into document embedding generation. CoRR abs/1607.05368 (2016)
9. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. CoRR abs/1405.4053 (2014). http://arxiv.org/abs/1405.4053
10. Ma, L., Lu, Z., Shang, L., Li, H.: Multimodal convolutional neural networks for matching image and sentence. In: IEEE International Conference on Computer Vision (ICCV) (2015)
11. van der Maaten, L., Hinton, G.: Visualizing high-dimensional data using t-SNE. J. Mach. Learn. Res. **9**, 2579–2605 (2008)

12. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. CoRR abs/1310.4546 (2013). http://arxiv.org/abs/1310.4546
13. Park, G., Im, W.: Image-text multi-modal representation learning by adversarial backpropagation. CoRR abs/1612.08354 (2016)
14. Pertusa, A., Gallego, A.-J., Bernabeu, M.: MirBot: a multimodal interactive image retrieval system. In: Sanches, J.M., Micó, L., Cardoso, J.S. (eds.) IbPRIA 2013. LNCS, vol. 7887, pp. 197–204. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38628-2_23
15. Pertusa, A., Gallego, A.J., Bernabeu, M.: MirBot: a collaborative object recognition system for smartphones using convolutional neural networks. Neurocomputing **293**, 87–99 (2018)