Deep neural networks and transfer learning applied to multimedia web mining

Conference Paper *in* Advances in Intelligent Systems and Computing · June 2018

DOI: 10.1007/978-3-319-62410-5_15

CITATIONS
11

READS
331

3 authors:

Daniel López-Sánchez
Universidad de Salamanca
23 PUBLICATIONS **45** CITATIONS
SEE PROFILE

Angélica González
Universidad de Salamanca
73 PUBLICATIONS **212** CITATIONS
SEE PROFILE

Juan Manuel Corchado Rodríguez
Universidad de Salamanca
941 PUBLICATIONS **14,788** CITATIONS
SEE PROFILE

Deep neural networks and transfer learning applied to multimedia web mining

Daniel López-Sánchez, Angélica González Arrieta, Juan M. Corchado

Department of Computer Science and Automation, University of Salamanca, Spain

Abstract. The growth in the amount of multimedia content available online supposes a challenge for search and recommender systems. This information in the form of visual elements is of great value to a variety of web mining tasks; however, the mining of these resources is a difficult task due to the complexity and variability of the images. In this paper, we propose applying a deep learning model to the problem of web categorization. In addition, we make use of a technique known as transfer or inductive learning to drastically reduce the computational cost of the training phase. Finally, we report experimental results on the effectiveness of the proposed method using different classification methods and features from various depths of the deep model.

Keywords: Web mining, deep learning, transfer learning

1 Introduction

During the last decades, the number of web pages available on the internet has grown exponentially. The proliferation of blog-hosting and free content management systems (CMS) such as WordPress, Blogger or Tumblr have contributed to this growth by making it possible for users with no experience in managing digital systems to share a variety of contents. The nature of these hosting services promotes the publishing of multimedia contents, especially images and videos. However, this democratization of the internet has originated new and challenging problems. Specifically, it has become increasingly difficult for users to find the content that they demand while avoiding related but undesired results. In addition, the availability of abundant multimedia content supposes challenge to recommender and search systems. Conventional recommender and search systems do not usually take into account the discriminative information provided by multimedia content, limiting themselves to the analysis of textual information. This is because mining multimedia data is a highly complex problem that frequently demands intensive computations and large training datasets.

On the other hand, the field of artificial vision, and specifically the sub-field of visual object recognition, has experienced a major breakthrough after the general adoption of the deep learning paradigm in recent years [10]. Deep learning models are composed of a previously intractable number of processing layers, which allows the models to learn more complex representations of the data by taking into consideration multiple levels of abstraction. This eventually led to a dramatic improvement in the state-of-the-art of visual object recognition, object detection and other related domains [4]. The keys

adfa, p. 1, 2011.

to success for deep learning are the complexity of the models and the availability of large datasets with training data. The major drawback of deep learning techniques is their computational cost both in training and test phases.

In this paper, we propose applying a state-of-the-art model in visual object recognition to the field of multimedia web mining. Specifically, we propose using a deep convolutional neural network (DCNN) to the task of web page categorization based on the available multimedia content. To overcome the problems of computational cost and the need for large training datasets, we propose using a technique called transfer learning, which makes it possible to use the knowledge gained while solving one problem to another different but somehow related problem.

2 Proposed method

In this section the proposed method is described in detail. First, the global structure and elements of the processing pipeline are presented, followed by a detailed description of each element in the chain. Our system is designed to perform the following task: given an URL, the system must (1) access that URL, extract all the images available in the web page, and filter those that do not contain discriminative information; (2) extract a feature descriptor from each image such that the classification problem becomes easier over that feature space; and (3) analyze each feature descriptor and combine the results to emit a prediction concerning the category of the web page. This process is schematically shown in Figure 1.

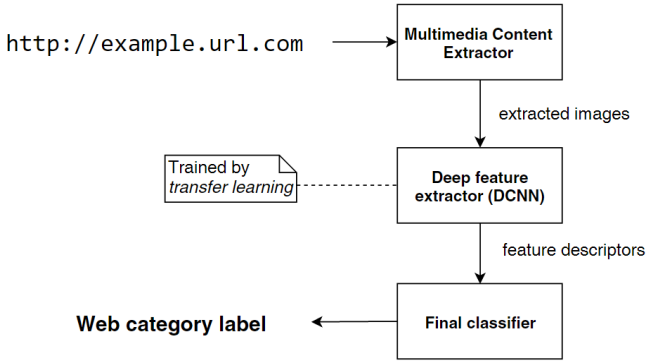

Fig. 1. Steps of the proposed method

2.1 Multimedia content extractor

The first module of the proposed method is in charge of extracting all the images that are present in a given web page and filter the ones that do not contain useful information (i.e. advertisements, banners, etc.). To do this, the system begins by downloading the HTML document to which the provided URL points. Then, we use the Beautiful Soup library [7] to analyze the structure and the hierarchy in the document. After this, the web page is represented as a tree whose leaves are the elements of the document (e.g.

titles, links, images, etc.). This representation of the document is explored exhaustively in order to find image elements; the URLs that point to those images are stored and later used to download the pictures.Several criteria can be applied to filter the extracted images; for example, it is possible to discard the images that contain a specific group of keywords in the alt attribute (e.g. we might discard images that contain the word "advertisement" in its alt attribute). Another possible approach consists of rejecting the images whose dimensions are outside a specific range. This is because images with rare proportions do not usually contain discriminative information. For example, very small images might correspond to navigation icons, and images of elongated shape tend to be advertisement banners.

2.2 Deep feature extractor

Once a number of images have been extracted from a web page, it would be possible to directly apply any classification method. However, the complexity of the image recognition problem that we are trying to solve demands a large number of training instances and a very complex model that can manage the difficulty of the classification problem. This is mainly due to the high intra-class variability of the samples from such artificial vision problems. Collecting such an extensive dataset can be a tedious and very time-consuming task. In addition, the training time of such a classifier would be very long even if we used complex computation parallelization techniques and expensive devices.

To overcome these limitations, we propose applying a technique known as transfer learning (see [8] for a recent survey on the topic). The key idea of this method is to apply the knowledge gained while solving one problem to a different but related problem. In practice, this technique has been mainly applied in the context of artificial neural networks. Here, some of the layers of the network are initialized with the weights learned by another network that was trained to solve a different problem; the remaining weights are initialized at random (as usual). A short training phase is then executed to tune the weights of the network that were randomly initialized. While the transferred layers may also be fine-tuned by executing various iterations of the chosen training algorithm, in principle this step is not mandatory. Avoiding this adjustment process will allow us to use any kind of classifier on top of the transferred neural layers, even if it is not compatible with the backpropagation method.

In the case study for this paper, we propose using a pre-trained model of DCNN. Specifically, the selected model is the VGG-16 DCNN [11] developed by the Visual Geometry Group at the University of Oxford[1]. The VGG-16 network achieved second place in the classification task of the Large Scale Visual Recognition Challenge 2014[2].

---

[1] The VGG16 model is available at the Visual Geometry Group web page: http://www.robots.ox.ac.uk/~vgg/research/very_deep/

[2] Large Scale Visual Recognition Challenge 2014 web page: http://www.image-net.org/challenges/LSVRC/2014/

It was trained on the ImageNet dataset, which consists of 14 million images belonging to 1000 categories; the model achieves 92.7% top-5 test accuracy on this benchmark. Although several models have outperformed VGG16, this model remains competitive with the state-of-the-art. In addition, it was designed with computational costs in mind. The number of parameters of the network was significantly reduced by using small 3×3 kernels in the convolution layers of the network. We chose the VGG16 model because it maintains a balance between computational costs and accuracy. Figure 2 shows the overall architecture of the network; we will refer to this figure to name each of the layers of the model in the experimental results section.

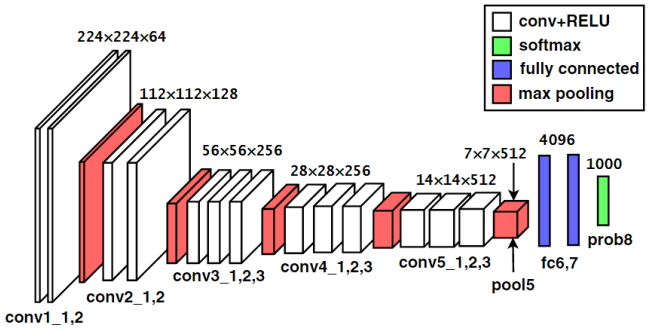

Fig. 2. VGG16 model architecture

In the proposed framework, the network is cut at a specific layer and the activations of the neurons in that layer are used as a representation of the input image. A simpler classifier is then applied over that feature space.

It has been proved that the outermost layers provide a more abstract and compact representation of the input images [12]. However, the final layers are more task-dependent and might not be useful if the target problem we want to address is very different from the problem the network was originally trained to solve. The layer that produces the most suitable representation for a given task must be determined empirically. To this end, the accuracy rates obtained with features from different layers and various classification methods are reported in section 3.

2.3 Final classifier

After the high level features have been extracted from the images using the deep neural network, a simpler classifier is in charge of emitting the final class prediction. Given that the features are sufficiently abstract, the use of a linear classifier is adequate. However, we also evaluated a simple nonparametric classifier that is not strictly a liner classifier, namely the k-Nearest Neighbor (kNN) algorithm.

In addition, the following linear classifiers were applied:

1. Support Vector Machine (C-SVM) with a linear kernel, as implemented in LIBSVM [1]. The decision function of this classifier is the following:

$$h(x) = sgn(w^T \varphi(x) + b) = sgn(w^T x + b)$$

Where *w* is calculated by solving the optimization problem presented in [2]. The multi-class support is obtained following a one-vs-one scheme.

2. Perceptron with linear activation function as implemented by *scikit-learn* [9]

$$h(x) = w^T x + b$$

Where *w* was optimized according to the mean squared error (mse) loss function. The stopping criteria was five iterations in all the experiments.

3. Logistic Regression (LR) as implemented in LIBLINEAR [3]. The decision function of this method is the following:

$$h(x) = \frac{1}{1 + e^{-\mu}} \; where \; \mu = w^T x + b$$

Here, the multi-class support was obtained using a one-vs-rest scheme. Note that despite the nonlinearity of the decision function, the decision boundary $\{x: h(x) = 0.5\}$ is a hyperplane and therefore this classifier is considered to be linear.

The majority of the web pages that we analyzed contained several images with relevant information. All the images must be taken into account to emit a prediction about the category of the web page. The simplest approach is taken where the final prediction is the most common label among the images of the web; when a tie occurs it is solved at random.

3 Experimental results

In this section, we evaluate the proposed method on a real word dataset for web page categorization. We provide insight into the suitability of transferred DCNN features by means of data visualization techniques and evaluate our system against individual image classification and web page categorization.

3.1 The dataset

To the best of our knowledge, there is no standard dataset for web page classification that focuses on visual content. For this reason we decided to collect a new database of web pages and the visual content present on them at the time. Our dataset consists of the images extracted from 75 web sites, uniformly distributed among five categories, namely “food and cooking”, “interior design”, “pets and wildlife”, “motor and races” and “fashion”. A total of 1,232 images were extracted from those web pages. The train/test split was arranged with 15 web pages for training and 60 web pages for testing purposes. As proved by our experimental results, the use of the transfer learning technique provides high accuracy rates in spite of the lack of a large training dataset.

Fig. 3. Class overlapping (interior design/pets and wildlife)

As expected, the artificial vision problem that emerged was of significant complexity. The dataset contains a high intra-class variability and some class overlapping (as shown in Figure 3).

3.2 DCNN feature visualization

In this section, we try to provide some insight into the suitability of the features at different depths of the network to solve the proposed problem. To do this, we use the t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm [5]. This method is a powerful nonlinear dimensionality reduction algorithm and is generally used to visualize high-dimensional data. It works by iteratively minimizing a non-convex cost function so that similar objects are modeled by nearby points and dissimilar objects are modeled by distant points. Due to the non-convexity of the cost function that is optimized with gradient descent, one may obtain slightly different results each time the algorithm is executed. If a given layer of the network produces a suitable representation, the activations of the neurons in that layer (obtained by feeding the network with the samples of our dataset) will form independent and easily separable clusters in the t-SNE 2D representation [5]. Figure 4 shows the 2D t-SNE embedding of the feature maps at different depths of the VGG16 model for all the images extracted from the collected web pages. Due to the similarity between the original problem VGG16 was trained to solve and the problem being addressed here, the features at deeper stages of the network produce t-SNE clusters with lower overlapping between classes. This suggests that features with a higher level of abstraction will propitiate better classification results.

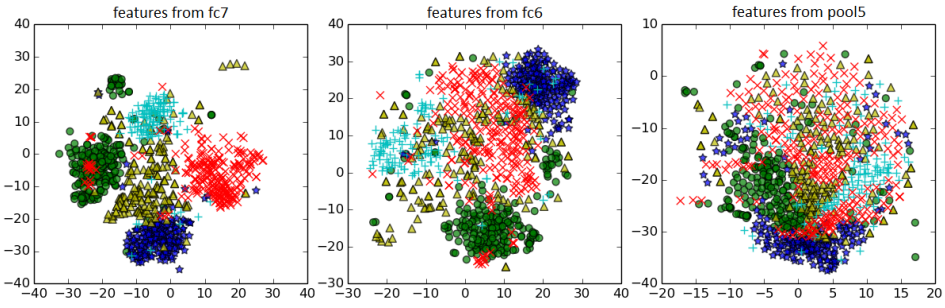

Fig. 4. t-SNE embedding of features at different depths. Food (●), motor (⋆), interior design (▲), animals (+) and fashion (×)

3.3 Classification accuracy results

In this section we report experimental results concerning the classification accuracy of both individual images and complete web pages. As described before, several classification methods were trained on the images extracted from 15 web pages, using the features at various depth levels of the network. The classification accuracy was then evaluated on the images extracted from the 60 remaining web sites. The accuracies for single image categorization are reported in Table 1. As explained before, the predictions for individual images are combined to categorize the complete web page. The accuracy rates for web page classification are shown in Table 2.

Table 1. Accuracy rates (%) on individual images

Features	SVM	LR	Perceptron	kNN
fc7	0.905	0.821	0.886	0.589
fc6	0.884	0.884	0.876	0.340
pool5	0.869	0.843	0.864	0.263
pool4	0.753	0.727	0.729	0.367
pool3	0.650	0.605	0.649	0.261
pool2	0.554	0.557	0.544	0.203

Table 2. Accuracy rates (%) on web categorization

Features	SVM	LR	Perceptron	kNN
fc7	0.966	0.9	0.93	0.56
fc6	0.916	0.95	0.933	0.31
pool5	0.9	0.85	0.916	0.25
pool4	0.816	0.816	0.766	0.4
pool3	0.766	0.583	0.766	0.25
pool2	0.733	0.55	0.7	0.2

4 Discussion and future work

In this paper, a novel framework for web page categorization was proposed. The system is able to classify web pages based on their visual content rather than their textual information. This makes the proposed technique very appropriate to modern web site analysis where visual elements have a dominant role. The major contribution of this paper is the application of transfer learning techniques to the problem of web page categorization. Experimental results show that this approach enables the construction of very accurate classifiers even when the artificial vision task to solve is significantly complex. In addition, our experiments show that competitive accuracy rates can be obtained with training phases of minutes, while training a complete deep neural network model to solve such a complex vision task typically requires hours or days, even if expensive specialized hardware is available. The second major advantage of the transfer learning approach is that it allows decent accuracy rates even if the training set is of

reduced length. The proposed approach could be further improved. First, a more sophisticated method to filter advertisements and other non-relevant content from web pages could be implemented. In addition, if more training data were available, it would be possible to adjust the weights of the deep neural network by executing various iterations of backpropagation. This would likely improve the accuracy rate of the proposed framework. Finally, a more advanced way of combining individual image predictions to categorize web pages could be developed, including techniques such as ensemble classification and mixture of experts [6].

Acknowledgements. The research of Daniel López-Sánchez has been financed by the University Faculty Training (FPU) program, reference number FPU15/02339.

5 References

[1] Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011). Software available at http://www.csie.ntu.edu.tw/~cjlin/

[2] Cortes, C., Vapnik, V.: Support-vector networks. Machine learning 20(3), 273–297 (1995)

[3] Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research 9, 1871–1874 (2008)

[4] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

[5] Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9 (Nov), 2579-2605.

[6] Masoudnia, S., & Ebrahimpour, R. (2014). Mixture of experts: a literature survey. Artificial Intelligence Review, 42(2), 275-293.

[7] Nair, V. G. (2014). Getting Started with Beautiful Soup. Packt Publishing Ltd.

[8] Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10), 1345-1359.

[9] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830.

[10] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Berg, A. C. (2015). Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115(3), 211-252.

[11] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[12] Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks?. In Advances in neural information processing systems (pp. 3320-3328).