



www.dream-go.ipp.pt

Real-time demand response and intelligent direct load control

Second DREAM-GO Workshop

University of Salamanca, Salamanca, Spain, March 22-23, 2017

Web Page Classification with Pre-Trained Deep Convolutional Neural Networks

Daniel López, Angélica González Arrieta, Juan M. Corchado

University of Salamanca, BISITE Research Group, Edificio I+D+i, C/ Espejo s/n, 37007 Salamanca, Spain

Abstract

In this paper, we propose mining the growing amount of information present on the internet in the form of visual content. We address the problem of web page categorization based on the multimedia elements present on it. To achieve this, our framework leverages a pre-trained deep convolutional neural network model, which is used as a feature extractor for later classification. This paper presents experimental results concerning the effectiveness of different classifiers trained with features extracted at various depths of the convolutional neural network.

Keywords: web mining, neural networks, deep 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 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 Framework

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.

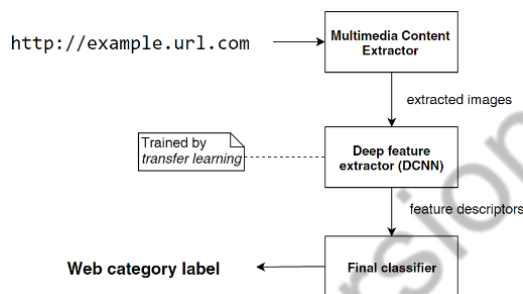


Fig. 1: Steps of the proposed framework

2.1 Visual Content Extraction

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 BeautifulSoup 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 neutral 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. The VGG-16 network achieved second place in the classification task of the Large Scale Visual Recognition Challenge 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.

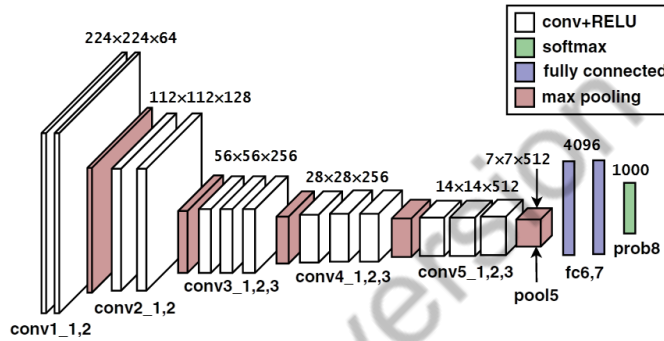


Fig. 2: VGG16 CNN architecture

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 Classifiers

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 linear 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:

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

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:

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

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.

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.

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
<i>fc7</i>	0.905	0.821	0.886	0.589
<i>fc6</i>	0.884	0.884	0.876	0.340
<i>pool5</i>	0.869	0.843	0.864	0.263
<i>pool4</i>	0.753	0.727	0.729	0.367
<i>pool3</i>	0.650	0.605	0.649	0.261
<i>pool2</i>	0.554	0.557	0.544	0.203

Table 2. Accuracy rates (%) on web categorization

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

4. Conclusions

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. This work has been supported by the European Commission H2020 MSCA-RISE-2014: Marie Skłodowska-Curie project DREAM-GO Enabling Demand Response for short and real-time Efficient And Market Based Smart Grid Operation - An intelligent and real-time simulation approach ref 641794.

References

- [1] Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10), P10008.
- [2] Bostock, M., Ogievetsky, V., & Heer, J. (2011). D³ data-driven documents. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12), 2301-2309.
- [3] Klusch, M. (Ed.). (2012). *Intelligent information agents: agent-based information discovery and management on the Internet*. Springer Science & Business Media.
- [4] Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1), 1-167.
- [5] McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, 415-444.
- [6] Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., & Bhattacharjee, B. (2007, October). Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement* (pp. 29-42). ACM.
- [7] Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23), 8577-8582.
- [8] Nguyen, D., Demeester, T., Trieschnigg, D., & Hiemstra, D. (2012, October). Federated search in the wild: the combined power of over a hundred search engines. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 1874-1878). ACM.
- [9] Pfalzner, S., & Gibbon, P. (2005). *Many-body tree methods in physics*. Cambridge University Press.
- [10] Schrenk, M. (2012). *Webbots, spiders, and screen scrapers: A guide to developing Internet agents with PHP/CURL*. No Starch Press.
- [11] Stefanidis, A., Crooks, A., & Radzikowski, J. (2013). Harvesting ambient geospatial information from social media feeds. *GeoJournal*, 78(2), 319-338.
- [12] Tapscott, D. (2008). *Grown Up Digital: How the Net Generation is Changing Your World HC*. McGraw-Hill.
- [13] Westerman, D., Spence, P. R., & Van Der Heide, B. (2014). Social media as information source: Recency of updates and credibility of information. *Journal of Computer-Mediated Communication*, 19(2), 171-183.
- [14] Sanchez Martin, A. J., de la Prieta Pintado, F., & De Gasperis, G. (2014, July). Fixing and evaluating texts: Mixed text reconstruction method for data fusion environments. In *Information Fusion (FUSION), 2014 17th International Conference on* (pp. 1-6). IEEE.