proceedings

MDPI

Proceedings

On the Effectiveness of Convolutional Autoencoders on Image-Based Personalized Recommender Systems †

Eva Blanco-Mallo [1,*], Beatriz Remeseiro [2], Verónica Bolón-Canedo [1] and Amparo Alonso-Betanzos [1]

[1] Campus de Elviña s/n, Universidade da Coruña, CITIC, 15071 A Coruña, Spain; vbolon@udc.es (V.B.-C.); amparo.alonso.betanzos@udc.es (A.A.-B.)

[2] Campus de Gijón s/n, Universidad de Oviedo, 33203 Gijón, Spain; bremeseiro@uniovi.es

* Correspondence: eva.blanco@udc.es

† Presented at the 3rd XoveTIC Conference, A Coruña, Spain, 8–9 October 2020.

Published: 19 August 2020

Abstract: Over the years, the success of recommender systems has become remarkable. Due to the massive arrival of options that a consumer can have at his/her reach, a collaborative environment was generated, where users from all over the world seek and share their opinions based on all types of products. Specifically, millions of images tagged with users' tastes are available on the web. Therefore, the application of deep learning techniques to solve these types of tasks has become a key issue, and there is a growing interest in the use of images to solve them, particularly through feature extraction. This work explores the potential of using only images as sources of information for modeling users' tastes and proposes a method to provide gastronomic recommendations based on them. To achieve this, we focus on the pre-processing and encoding of the images, proposing the use of a pre-trained convolutional autoencoder as feature extractor. We compare our method with the standard approach of using convolutional neural networks and study the effect of applying transfer learning, reflecting how it is better to use only the specific knowledge of the target domain in this case, even if fewer examples are available.

Keywords: personalized recommendation; image-based recommendation system; feature extraction; convolutional autoencoder; convolutional neural network; data augmentation

1. Introduction

With the advent of e-commerce, social networks dedicated to sharing product reviews began to become popular, leading to the integration of different personalized recommender systems (RS) with great success on various on-line platforms. Starting from the premise that a picture is worth more than a thousand words [1], our goal is to make use of the data available in TripAdvisor to build a personalized gastronomic image-based RS. One of the first attempts to use visual information in personalized RS was proposed by He et al. [2], using a convolutional neural network (CNN) to extract the deep features of product images that are then processed through matrix factorization. As for the use of TripAdvisor data, Zhang et al. [3,4] also use them for recommendation purposes, but do not consider the visual information at all. A more related approach to ours is the one presented by Díez et al. [1], who also use TripAdvisor images but to determine authorship of the images, aiming at providing explainable recommendations. To the best of our knowledge, our work is the first one that studies the effect of using only images to characterize user tastes in a personalized recommendation system. Unlike existing approaches, we propose to use a convolutional autoencoder (CAE) as a feature extractor, because (1) it works better than standard approaches based on pre-trained CNN, and (2) it is less computationally expensive.

2. Materials and Methods

The problem addressed in this manuscript is defined as a binary classification task in which we have some triads with either one of two labels, $(u, r, i) \rightsquigarrow 0|1$, where u is a user who took the image i of a restaurant r he/she visited. As for the two labels, 0 means that the user u does not like the restaurant r reflected in i, while 1 stands for the opposite. A network that learns on triads of users and photos (u, r, i) is proposed to provide a personalized image-based recommendation. Users and restaurants are represented by a one-hot codification and then mapped into 512-dimensional embeddings, while photos are codified by a CAE. The CAE, based on the one proposed by Chollet [5], was trained using the entire set of images available, and then used to encode the images used at the model input. This code is processed by a fully connected (FC) layer, and then concatenated with the one-hot embeddings, resulting in a single vector of 1536 features. Next, a series of FC, ReLU and dropout layers follow, ending with the sigmoid activation function that generates a probability value, where 0 means that the user u does not like the restaurant r depicted by i, and 1 that he/she likes it.

The data used for experimentation purposes were collected in 2018–2019 from TripAdvisor reviews of restaurants in three cities of different sizes [1]: Santiago de Compostela (SGC), Barcelona (BCN) and New York (NYC). The original dataset was divided to obtain the training and test sets, following this procedure: for each user, reviews are separated into positive and negative, with 1 review (if there is more than one) for the test set and $N - 1$ for the training set; once this procedure is completed, if there is any review in the test set that belongs to a restaurant that is not included in the train set, then all its images are moved to the train set. The idea is to guarantee that all the users and restaurants evaluated in the test set are also in the train set. To select the hyper-parameters of the proposed architecture, a validation set was obtained applying the above procedure to the training set. Due to the high imbalance of the data, oversampling is applied on the minority class until an approximate balancing ratio of 1 is reached.

3. Results

All the experiments were performed on a computer equipped with a GeForce Titan XP 12GB GPU from NVIDIA, an Intel Core i7-4790 CPU @ 3.60GHz x 8, and 16 GiB memory. The implementation of the model and baseline methods is in Keras (https://keras.io/), using the Adam as optimizer and the HeUniform for weight initialization. The training process was carried out by setting a batch size of 32, a patience of 12 and a maximum of 100 epochs. The outputs were monitored by using the balanced score (B-score) metric, which represents the harmonic mean of sensitivity and specificity: B-score $=$ 2 $*$ ((sensitivity * specificity) / (sensitivity + specificity)). The effectiveness of the CAE as a feature extractor is contrasted with a CNN, considered a benchmark in supervised image classification. Specifically, ResNet50 [6] with weights pre-trained on ImageNet is used, with and without parameter fine-tuning. In an RS, not only is it important to recommend those items that the user likes, but also not to recommend those items that the user does not like. Thus, we focus on sensitivity and specificity metrics and propose to use the B-score previously defined.

In light of the results shown in Table 1, the best performance is achieved by the CAE, specifically regarding specificity. The only city where the ResNet50 achieves a higher specificity is SGC and, even so, the balance between the classification of both classes is still lower than that obtained with the CAE. In general, terms, the worst scenario occurs when the ResNet50 is used without parameter fine-tuning, since it was trained on ImageNet and it is not adjusted to the problem at hand. Therefore, it is important to emphasize that the use of transfer learning was not as useful as intended and better results are obtained by using only specific knowledge of the target domain. In relation to the time per epoch, the CAE stands out without a doubt, requiring less than half the time that ResNet50 does and even less than nine times less if parameter fine-tuning is applied. Hence, our approach is more cost-effective not only in terms of performance, but also of processing time.

Table 1. Comparison of results using three different techniques for the image encoding step: using a CAE, a CNN pre-trained on ImageNet and a fine-tuned CNN (CNN_FT).

	Sensitivity			**Specificity**			**B-Score**		
	CAE	**CNN**	**CNN_FT**	**CAE**	**CNN**	**CNN_FT**	**CAE**	**CNN**	**CNN_FT**
NYC	0.7454	0.7261	0.7933	0.7594	0.7071	0.6724	0.7523	0.7165	0.7279
BCN	0.6175	0.8546	0.6738	0.8006	0.4480	0.6176	0.6972	0.5878	0.6445
SGC	0.7629	0.8453	0.7318	0.7905	0.6047	0.8181	0.7765	0.7050	0.7725

4. Conclusions

In this work, we explore the potential of modeling both users and restaurants using images as single source, demonstrating that the use of CAEs rather than CNNs is more convenient, not only considering performance but also resources. Taking into account that the high imbalance of classes and the few examples per user are characteristics in the context of the recommendations, it is interesting to further investigate possible transfer learning approaches, other than those based on parameter transfer, that may be more suitable in these scenarios. Our future research involves the integration of the proposed method into an RS that takes into account additional information, and its application to other RS that already deal with images.

Funding: This research has been financially supported in part by European Union FEDER funds, by the Spanish Ministerio de Economía y Competitividad (research project PID2019-109238GB), by the Consellería de Industria of the Xunta de Galicia (research project GRC2014/035), and by the Principado de Asturias Regional Government (research project IDI-2018-000176). CITIC as a Research Center of the Galician University System is financed by the Consellería de Educación, Universidades e Formación Profesional (Xunta de Galicia) through the ERDF (80%), Operational Programme ERDF Galicia 2014–2020 and the remaining 20% by the Secretaria Xeral de Universidades (ref. ED431G 2019/01).

Acknowledgments: Thanks to NVIDIA Corporation for donating the Titan Xp GPUs used in this research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Díez, J.; Pérez-Núñez, P.; Luaces, O.; Remeseiro, B.; Bahamonde, A. Towards explainable personalized recommendations by learning from users' photos. *Inf. Sci.* **2020**, *520*, 416–430.
2. He, R.; McAuley, J. VBPR: visual bayesian personalized ranking from implicit feedback. In Prcoceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; pp. 144–150.
3. Zhang, H.; Ji, P.; Wang, J.; Chen, X. A novel decision support model for satisfactory restaurants utilizing social information: A case study of TripAdvisor.com. *Tour. Manag.* **2017**, *59*, 281–297.
4. Zhang, C.; Zhang, H.; Wang, J. Personalized restaurant recommendation method combining group correlations and customer preferences. *Inf. Sci.* **2018**, *454*, 128–143.
5. Chollet, F. Building Autoencoders in Keras. *The Keras Blog*; **2016**.
6. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Prcoceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 30 June 2016; pp. 770–778.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).