Author Profiling From Images Using 3D Convolutional Neural Networks*

Eduardo Valdez-Rodríguez, Hiram Calvo, and Edgardo Felipe-Riverón

Centro de Investigación en Computación, Instituto Politécnico Nacional
J.D. Bátiz e/ M.O. de Mendizábal, 07738, Ciudad de México, México
jvaldezr1000@alumno.ipn.mx, hcalvo@cic.ipn.mx, edgardo@cic.ipn.mx

Abstract. With this work we participate in the competition on the author profiling task on the MEX-A3T track at IberEval 2019. Author profiling task aims to identify gender, occupation and location from images or text of Mexican Twitter users. We propose a 3D Convolutional Neural Network for solving this task, using visual information, in this case images extracted from the user's profile which are grouped to create a unique input of each user.

Keywords: Author profiling · 3D CNN · Visual information

1 Introduction

Author Profiling (AP) consists on identifying demographic information such as age, gender, location, occupation, native language, personality traits, etc. [8], to obtain a description from a particular user. This information can be obtained from several sources of information, in this case through texts and images from the user's Twitter profile due to the large quantity of information available in this medium.

In this work we focus on obtaining the profile of an author using images only; however, using only this source of information is a challenging task due to the interaction of users in social media, for example a group of friends can share the same nature of the image on their Twitter account. Another issue is related to gender equality, making gender identification complicated because now a man or woman can share the same images. That is why in this work we propose to work with groups of images instead of single images from a user's profile. With this group of images we aim to capture the history of each user's profile in such a way that it will not be of importance if another user uses the same image; to classify these groups of images we propose 3D Convolutional Neural Network (CNN) models capable of extracting features from these groups of images.

This work is structured as follows: in Section 2 we describe the previous work related to this task; in Section 3 we describe our proposal; in Section 4 we show the results obtained from this work, and finally in Section 5 we draw our conclusions.

2 Previous work

In this work we focus on Mexican Twitter users and on three classes of demographic information: **gender**, **location** and **occupation**; most of the previous works on AP have been devoted to the analysis of textual information, disregarding information from other modalities, such as visual information, that could be potentially useful for improving the performance of AP methods.

Some of the first related works were devoted to obtainin gender using textual information are developed by Argamon et al. [3][4]. They proposed combinations of textual attributes ranging from lexical features such as content words and function words. Another work by Ortega et al. [12], uses syntactical features such as personal phrases to detect gender on texts.

Merler et al. [11] uses images extracted from the user profile and categorize each one with independent classifiers, to obtain gender from those categories. Taniguchi et al. [13] categorizes

* We thank Instituto Politécnico Nacional (SIP, COFAA, EDD, EDI and BEIFI), and CONACyT for their support for this research.

each image of the profile using CNNs and then performs classification on those categories to predict gender. Although Merler et al. [11] and Taniguchi et al. [13] work with images, they rely on an intermediate stage that recognizes certain features in images, and thus, an additional process of labeling and training is required. A similar work is made by Álvarez-Carmona et al. [1]; they use a multimodal approach to identify gender and age from Twitter users using both texts and single images. To our knowledge, there are no works devoted to author profiling using images directly, without depending on a previous categorization of images prior training. In the next sections we give a detailed description of our proposal.

3 Proposal

In this section we describe our method using visual information from Mexican Twitter users; we used 3D CNN models and images to identify the AP **gender**: male and female, **location**: north, northwest, northeast, south, southeast and center, and **occupation**: arts, student, social, sciences, administrative, health, sports and others.

It is worth to mention that we are working with Mexican Twitter users, so their native language is Spanish while all the previous mentioned work was focused on English language users. Another important point is that, as far as we know, there are no other works trying to predict **location** and **occupation** from both textual and visual information.

Finally, in this work we use only visual information provided by the users, that is, we will use images taken from their Twitter profiles. From the aforementioned works we take the idea that CNNs are capable of extracting features from single images and classify them in different classes. We propose a 3D CNN capable of classify gender, location and occupation at the same time from groups of images instead of single images.

3.1 Dataset

The dataset was provided by the AP task on the MEX-A3T at IberEval 2019 [2]. It consists of images and texts extracted from Mexican Twitter users to identify gender, occupation and location. For solving this task we use images only. The dataset contains 11 images per user (in few cases the number of images is less than 11 depending on the Twitter user), the full dataset consists of 3,500 users, resulting in 38,500 images approximately. In Fig. 1 we show samples of several images of this dataset.

Fig. 1. Sample images from a Twitter user.

By analyzing the images given by the users we found that using a single image as in Álvarez-Carmona et al. [1], determine all labels was not possible because some users, for example a man or woman, can share the same image. This is why we decided instead of using a single image, we could use a group of images shared by a user. As we are using a 3D CNN, we adapted the dataset grouping several images in a single volume for each user as shown in Fig. 2.

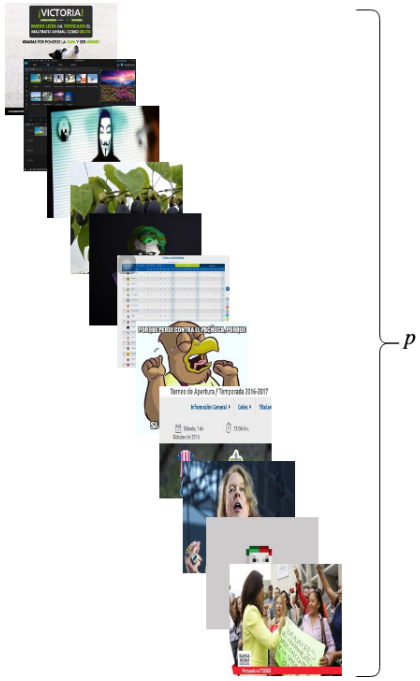

Fig. 2. Group of images from a Twitter user as input for the 3D CNN, where p is the number of images per user.

3.2 3D CNN

We use two 3D CNN models based on [5] and [6]; they use a 3D CNN to classify human actions on videos; this kind of CNNs work with temporal information such as videos; a video can be seen as a sequence of images in time. We do not posses that temporal relationship but we have groups of images that can be processed as sequences by the 3D CNN. The representation of a convolutional and fully connected layer is shown in Fig. 3.

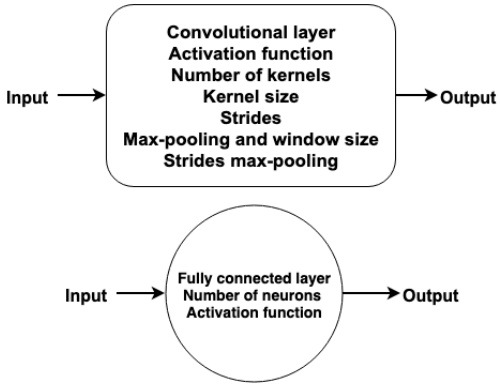

Fig. 3. Single layer representation of the convolutional layer and a fully connected layer.

The 3D CNN Model 1 is shown in Figure 4 and it is based on [6], where h is the height of the patch, w is the width of the patch, p is the number of images per user—in this case we decided to use 11 images per group— and n is the batch size. It consists of four convolutional layers to extract features; in this CNN we use 3D kernels of size 2x3x3 and Max-pooling with size 2x2x2; finally we add a classification stage.

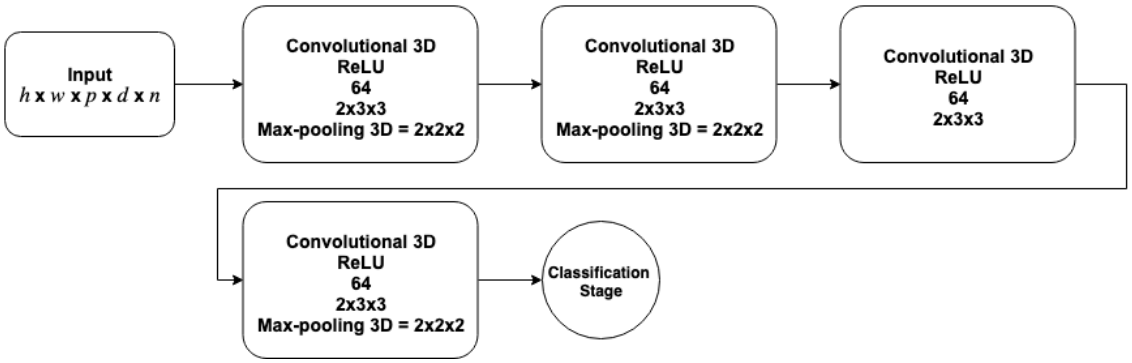

Fig. 4. 3D CNN Model 1.

The 3D CNN Model 2 is shown in Figure 5 and it is based on [5], where h is the height of the patch, w is the width of the patch, p is the number of images per user—in this case we decided to use 11 images per group— and n is the batch size. It consists of seven convolutional layers to extract features; in this CNN we use 3D kernels of size 2x3x3 and Max-pooling with size 2x2x2; finally we add a classification stage.

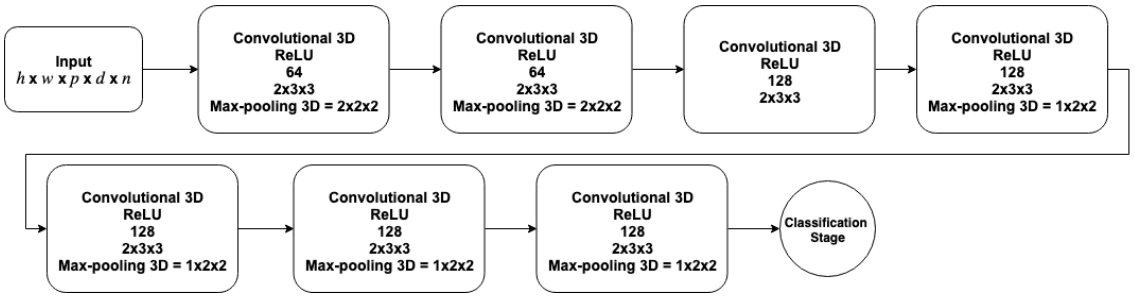

Fig. 5. 3D CNN Model 2.

The main differences between both models is the Max-pooling operation performed on the last layers of Model 2, this operation reduces the input volume just in h and w dimension to extract features at lower resolution. Also the number of layers used to extract features from the input volume is different in both models.

Classification stage. We needed to adapt the output of the models to classify each profiling setting separately; we use Fully Connected layers to perform the classification, all of them with ReLU as activation function; as loss function we use the Categorical Crossentropy function [9] given by Eq. 1.

$$L(y, \hat{y}) = -\sum_{j=0}^{M}\sum_{i=0}^{N}(y_{ij} * log(\hat{y_{ij}})) \tag{1}$$

where $\hat{y}$ is the predicted value and y the target value. Categorical crossentropy will compare the distribution of the predictions (the neurons in the output layer, one for each class) with the true distribution, where the probability of the true class is set to 1 and 0 for the other classes. This output layer consists of three stages, the first one classifies gender (male and female) using 2 output neurons, the second one classifies location (north, northwest, northeast, south, southeast and center) using 6 output neurons and finally the third one classifies occupation (arts, student, social, sciences, administrative, health, sports and others) using 8 output neurons. In Figure 6 we can observe how each stage is divided.

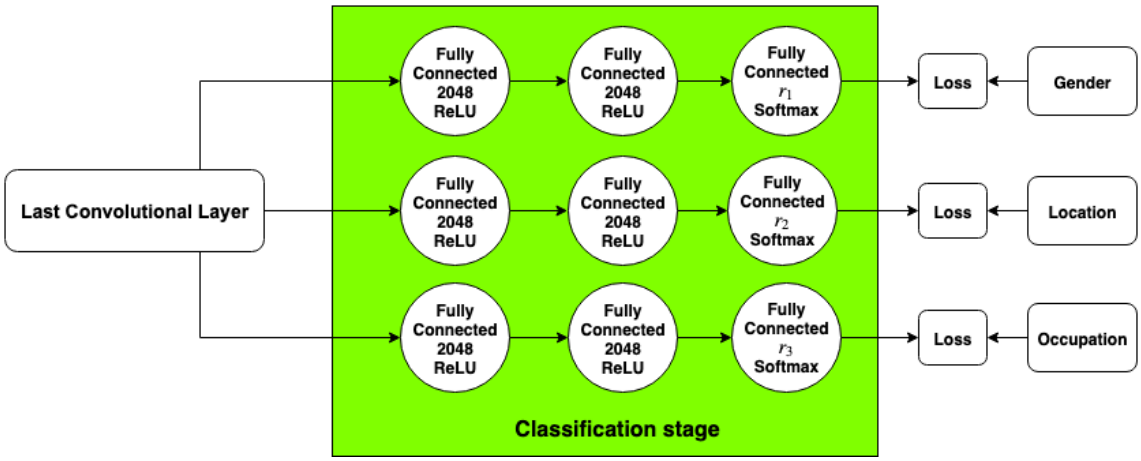

Fig. 6. Classification stage, where r is the number of output neurons for each profiling task and $r_1 = 2$, $r_2 = 6$ and $r_3 = 8$.

4 Experiments and results

The dataset provided by [2] consists of training and test sets; the training set consists of 3,500 groups of images per user and 11 images for each group and the testing set consists of 1,500 groups of images per user with the same number of images.

We trained both models using the training set and using the Backpropagation (BP) method [10] and ADAM optimizer [7] to update the weights during training, with learning rate equal to 0.05 and decay equal to 0.7; we used 100 iterations and a batch size of 13. We trained the model on two GPU NVIDIA GTX 1080Ti; it took five hours for training and less than a second for testing a single group of images.

As mentioned before we were participants of the AP task on the MEX-A3T at IberEval 2019; we compare our method with *CerpamidTeam* as their results were made public for this contest as well; they use the textual information to perform their prediction. From Tables 1 to 6 we report the evaluation of our 3D CNN models, and present the comparison with CerpamidTeam. We use the metrics given by the organizers of the contest such as F measure, Accuracy (Acc), Precision (P) and Recall (R).

Table 1. Comparison between our method and CerpamidTeam on gender.

Method	F(P, R)	Acc	P	R
CerpamidTeam 1	0.8366	0.8347	0.8412	0.8347
CerpamidTeam 2	0.8331	0.8313	0.8392	0.8313
3D CNN, Model 2	**0.5227**	**0.5227**	**0.5230**	**0.5227**
3D CNN, Model 1	0.4711	0.4840	0.4823	0.4840

Table 2. Comparison between our method and CerpamidTeam on location.

Method	F(P, R)	Acc	P	R
CerpamidTeam 1	0.5020	0.6307	0.5214	0.4934
CerpamidTeam 2	0.4844	0.6153	0.4988	0.4766
3D CNN, Model 1	**0.1482**	**0.2393**	0.1645	0.1573
3D CNN, Model 2	0.1098	0.1687	**0.1674**	**0.1846**

Table 3. Acc comparison between our method and CerpamidTeam on different values for location.

Method	center	southeast	northwest	north	northeast	west
CerpamidTeam 1	0.6889	0.3864	0.6667	0.1667	0.7091	0.2890
CerpamidTeam 2	0.7021	0.3984	0.6651	0.2535	0.7297	0.2634
3D CNN, Model 1	**0.4177**	**0.0430**	**0.0952**	0.0260	**0.2066**	**0.1008**
3D CNN, Model 2	0.3606	0.0000	0.0867	**0.0491**	0.1377	0.0250

Table 4. Comparison between our method and CerpamidTeam on occupation.

Method	F(P, R)	Acc	P	R
CerpamidTeam 1	0.3957	0.6587	0.4331	0.3744
CerpamidTeam 2	0.3824	0.6613	0.4201	0.3619
3D CNN, Model 1	**0.1196**	**0.2693**	0.1235	**0.1284**
3D CNN, Model 2	0.0951	0.2300	**0.1402**	0.1260

Table 5. Acc comparison between our method and CerpamidTeam on each occupation.

Method	others	arts	student	social	sciences	sports	admin	health
CerpamidTeam 1	0.1026	0.2514	0.8597	0.5695	0.1982	0.3077	0.5134	0.2564
CerpamidTeam 2	0.1333	0.3333	0.8648	0.5508	0.2075	0.3500	0.4785	0.2469
3D CNN, Model 1	0.0000	0.0902	**0.4495**	**0.1390**	**0.0654**	0.0000	**0.2131**	0.0000
3D CNN, Model 2	0.0000	**0.1181**	0.4394	0.0701	0.0408	0.0000	0.0926	0.0000

4.1 Discussion

As shown in previous tables we observe that predicting the AP of a Twitter user using only images is a difficult task due to the generality of purpose of images on this platform, and even if we look at the images is hard to classify them; nevertheless we obtain better results using the 3D CNN Model 2 on gender and using the 3D CNN Model 1 on location and occupation. As far a we know there are not other participants who solve the task using images, so we cannot compare with the methods using text. Also both models work with volumes of images instead single images as described in previous work. One disadvantage from our method is that we need the same number of images per user and if a user has less number of images we are not able to determine precisely his or her profile.

5 Conclusions and future work

AP using only visual information such as images from Twitter is a difficult task due to the ambiguity and the source of them, although we did not obtain high results, we developed a new framework in which we can obtain AP using only images. Also we created a framework in which we can use groups of images instead of single images, avoiding the problem of shared images between users; additionally, our models allow to classify simultaneously directly from the images three different aspects of the profile.

As a future work we plan to create a multimodal 3D CNN model which is able to combine visual and textual information.

References

1. Álvarez-Carmona, M.A., Pellegrin, L., Montes-y Gómez, M., Sánchez-Vega, F., Escalante, H.J., López-Monroy, A.P., Villaseñor-Pineda, L., Villatoro-Tello, E.: A visual approach for age and gender identification on Twitter. Journal of Intelligent & Fuzzy Systems **34**(5), 3133–3145 (2018)

2. Aragón, M.E., Álvarez-Carmona, M.Á., Montes-y Gómez, M., Escalante, H.J., Villaseñor-Pineda, L., Moctezuma, D.: Overview of MEX-A3T at IberLEF 2019: Authorship and aggressiveness analysis in mexican spanish tweets. In: Notebook Papers of 1st SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Bilbao, Spain, September (2019)
3. Argamon, S., Koppel, M., Fine, J., Shimoni, A.R.: Gender, genre, and writing style in formal written texts. Text-Interdiciplinary Journal for the Study of Discourse, **23**(3), 321–346 (2006)
4. Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Mining the blogosphere: Age, gender and the varieties of self-expression. First Monday **12**(9) (2007)
5. Hou, R., Chen, C., Shah, M.: Tube convolutional neural network (T-CNN) for action detection in videos. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5822–5831 (2017)
6. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**, 221–231 (2010)
7. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
8. Koppel, M., Argamon, S., Shimoni, A.R.: Automatically categorizing written texts by author gender. Literary and linguistic computing **17**(4), 401–412 (2002)
9. LeCun, Y., Boser, B.E., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W.E., Jackel, L.D.: Handwritten digit recognition with a back-propagation network. In: Advances in neural information processing systems. pp. 396–404 (1990)
10. LeCun, Y.A., Bottou, L., Orr, G.B., Müller, K.R.: Efficient backprop. In: Neural networks: Tricks of the trade, pp. 9–48. Springer (2012)
11. Merler, M., Cao, L., Smith, J.R.: You are what you tweet... pic! gender prediction based on semantic analysis of social media images. In: 2015 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2015)
12. Ortega-Mendoza, R.M., Franco-Arcega, A., López-Monroy, A.P., Montes-y Gómez, M.: I, me, mine: The role of personal phrases in author profiling. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 110–122. Springer (2016)
13. Taniguchi, T., Sakaki, S., Shigenaka, R., Tsuboshita, Y., Ohkuma, T.: A weighted combination of text and image classifiers for user gender inference. In: Proceedings of the Fourth Workshop on Vision and Language. pp. 87–93 (2015)