Using Convolutional Neural Networks for Content Extraction from Online Flyers

Alessandro Calefati, Ignazio Gallo, Alessandro Zamberletti, Lucia Noce
Universita' degli Studi dell'Insubria
Varese, Italy
http://artelab.dicom.uninsubria.it/

ABSTRACT

The rise of online shopping has hurt physical retailers, which struggle to persuade customers to buy products in physical stores rather than online. Marketing flyers are a great mean to increase the visibility of physical retailers, but the unstructured offers appearing in those documents cannot be easily compared with similar online deals, making it hard for a customer to understand whether it is more convenient to order a product online or to buy it from the physical shop. In this work we tackle this problem, introducing a content extraction algorithm that automatically extracts structured data from flyers. Unlike competing approaches that mainly focus on textual content or simply analyze font type, color and text positioning, we propose a new approach that uses Convolutional Neural Networks to classify words extracted from flyers typically used in marketing materials to attract the attention of readers towards specific deals. We obtained good results and a high language and genre independence.

CCS Concepts

•**Computing methodologies** → **Machine learning;** *Machine learning approaches; Machine learning algorithms;*

Keywords

Content Extraction; Portable Document Format; Convolutional Neural Network; Marketing Flyers.

1. INTRODUCTION

Although e-commerce has been increasing in popularity in recent years, unstructured documents such as marketing flyers and advertising emails are still effective ways to promote products and special offers. In this study we propose a novel content extraction algorithm to automatically extract entities of interest from marketing flyers containing commercial product offers.

Most of the deals appearing within commercial flyers refer to physical retailers, and the data that can be gathered from those marketing documents is particularly appealing to online price comparison shopping engines to fill the existing gap between online and physical shopping. In fact, a searchable collection containing both physical deals extracted from marketing flyers and offers gathered from online listings would allow customers to determine whether it is more convenient to order a product online and wait for order preparation, shipping and delivery, or to physically drive to the retailer location and buy it straight from the shelf. This represents a great opportunity for physical retailers to compete against online marketplaces.

The PDF standard is increasingly being used by physical retailers to create marketing materials that maintain the same visual characteristics on every different device. This is particularly important, as commercial flyers typically contain many graphic elements specifically designed to let customers quickly understand the positions of relevant offers within the pages or to attract their attention towards particular deals or sections. Recent works have shown that, when dealing with not much structured documents containing lots of graphic elements, textual features are not discriminative enough to accurately identify all the entities of interest. In fact, visual characteristics are typically used to highlight and categorize paragraphs of text, and therefore represent a large amount of information that can and should be used to more accurately distinguish entities of interest having low discriminative textual contents.

Casting the problem to PDF documents, existing studies transform the processed PDF content streams into raw text, wasting the visual formatting information contained within the documents, such as font type, size, color and text positioning. If we delve even deeper into the world of PDF documents and we focus on PDF marketing flyers, it has been proved that combining textual information with additional features describing the visual characteristics of the deals in the processed documents increases classification performances.

The goal of this study is to underline the important role that visual features have in distinguishing different entities of interest within highly unstructured but visually rich documents, considering images that contain them. Unlike other works in literature, we propose novel and simpler approach that uses images of words retrieved directly from flyers avoiding visual distortions that may arise when converting PDF to other formats such as HTML.

Although in this study we focus exclusively on marketing materials, the proposed approach is not handcrafted for that specific domain; as such, it can be used in a wide variety of

DocEng2016 September 13–16, 2016, Vienna, Austria

DOI: 10.1145/1235

contexts in which the visual characteristics of the processed documents are discriminative for the entities that need to be identified.

2. RELATED WORKS

In the following paragraphs we introduce some works similar to our paper.

Our study is analogue to the method proposed by Gallo *et al.* [4], the substancial difference involves the first phase: instead of extracting and analyzing textual and visual features from PDF, the novel approach that we are discussing, uses Convolutional Neural Networks, which actually represent the state-of-the-art in image classification.

The same task has been approached by Apostolova *et al.* [1] that proposed a method to retrieve information from PDF or HTML document using some textual and visual features. Our goal is the same of the one discussed in the paper, but we used different approach, not extracting information from files but using raw images gathered from PDF. As regards to the use of CNNs, Simard *et al.* [7] proposes a set of concrete best practices that document analysis researchers can use to get good results with this kind of neural networks. The paper suggests to enlarge dataset, adding some kind of distorted data beyond the available ones, and supplies a novel architecture of CNN that fits better to solve many problems usually encountered in document analysis. As described in this study, we expanded our dataset extracting a token five times rather than only one time before the CNN training.

Another paper which aims to enhance CNN's performance is the one proposed by Chellapilla *et al.* [2] that studied different strategies to speed up the use of CNNs. Discussed approaches are: unrolling the convolution, using of basic linear algebra subroutines and using the graphic processing unit (GPU) instead of CPU for calculation. These three methods are then compared to determine which one has greater impact on CNNs' performance, showing that using GPU the execution time is 4x faster. For our experiments we used computer equipped with two high performance video cards. So CNN calculus has been done on GPU breaking down training time at about 3 hours.

3. PROPOSED METHOD

The processing pipeline of the proposed approach is presented in this section: (i) single tokens are extracted from the processed marketing flyer; (ii) each token gathered is classified using a properly trained CNN classifier from AlexNet architecture [6]; (iii) neighbouring words having same classification result are merged into semantically correlated paragraphs; (iv) paragraphs representing correlated product titles, descriptions and prices are further merged together to identify the deals contained within the processed flyer. The whole pipeline is summarized in Fig. 1 and described in detail in the remainder of this section.

3.1 Token Classification and Aggregation

Token classification is carried out using CNN classifier trained to produce output of the following classes {*title*, *description*, *price*, *not-class*}. Before training the model, we extract tokens from the PDF flyer using PDFTextStripper, which is a java class of the PDFBox library. It extracts each word from the page with also additional informations like

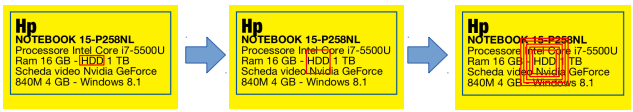

Figure 2: Pipeline of the extraction phase.

coordinates, useful to get the rectangle sourrounding the token drawn on the page. For our experiment, we used PDF files with readable text inside. Working with images saved as PDF, the proposed approach would not work, because PDFTextStripper would not extract any token. In this case the use of OCR algorithms could help to solve the problem. During training stage, for each token extracted from the ground-truth, we do a squarify operation, since almost tokens have rectangular shape. To transform a rectangle into a square, first, we calculate the difference between the longest and the shortest side, then we divide it by 2 and finally we stretch the shortest side adding the half of the difference between the width and the height of the token's rectangle. Then we magnify the extraction square five times and all page crops obtained are used to train the model. Algorithm involved is shown in Fig. 2. Magnification factor used to expand bounding boxes is not fixed, it is dependent on the ratio between the height of the entire page and the height of the token. The aim was to enlarge smaller squares more than bigger ones looking for other elements useful for the classification with CNN. An example of extracted token is shown in Fig. 3. Steps described above have been repeated in the classification phase. The membership of a token to a class is determined by number of votes received from the CNN. As proposed by [7], with this approach we got a dataset five times bigger than extracting single image for token.

Tokens extracted from the processed flyers are classified as belonging to one of the entities of interest listed in Table 2. The classification task is carried out using a CNN classifier from Caffe [5]. CNN classifiers have high accuracy when trained using a large amount of images, but they need a lot of training time and powerful hardware.

Once every token in the page has been assigned to a class of interest, they need to be aggregated to form products titles, descriptions and prices. This aggregation task is carried out using an ad-hoc clustering algorithm that takes into account the class and the distance from the surrounding elements of a token within the processed flyer.

At its first iteration, the algorithm selects the bounding box of a random seed token classified as belonging to either Title, Price or Description class, and tries to join that bounding box with all the other neighbouring bounding boxes of tokens classified as belonging to the same class c that are located at a distance $d < \epsilon$. This newly formed bounding box is then added to the page in place of all the joined bounding boxes. At each iteration a new seed token, that has not been previously selected, is chosen. The algorithm stops when all the tokens have been aggregated.

The result of this merging phase is a set of bounding boxes that represent titles, descriptions and prices of all the offers available on the flyer (see Fig. 4).

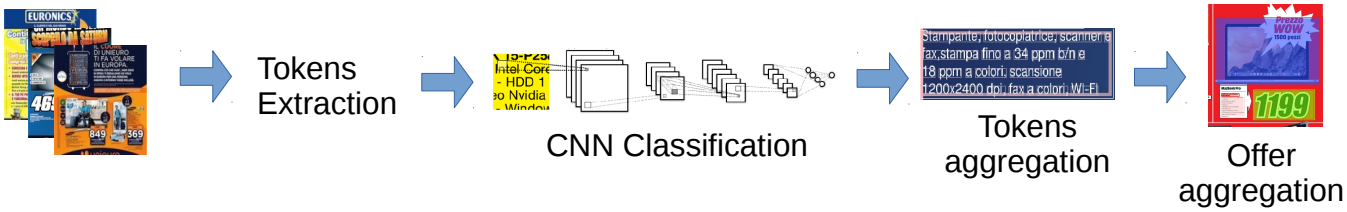

Figure 1: Pipeline of the proposed method.

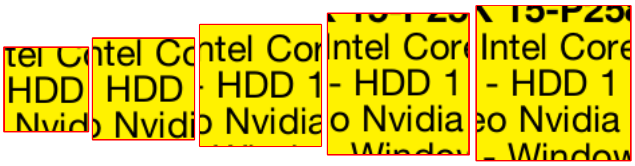

Figure 3: Example of squarified tokens then classified by CNN.

3.2 Offer Aggregation

The offers extraction process from each flyer is the last step of the presented method. This is not a trivial task as it cannot be carried out simply by considering the minimum distance between the various elements that form an offer. In fact, there are many cases in which one or more of the bounding boxes for the 3 relevant elements that make up an offer (Price, Title and Description) are visually closer to the bounding boxes of elements from another offer. In such cases, clustering exclusively on the basis of the distance between different bounding boxes does not lead to optimal results.

A better approach consists in clustering the bounding boxes in such way that the coverage provided by the final clusters over the processed page is maximized. This approach is motivated by the fact that marketing documents do not usually have many void areas, because retailers typically try to lower printing costs by adding as many offers as they can within each page to reduce the total size of the final flyer. As such, each offer within a page is usually localized in a particular area, and its bounding box has a minimal overlap with the other offers.

The textual information associated with an offer O is a triple (T, D, P) composed of a Title T, a Description D and a Price P. In this work, product images are not taken into account because finding the correct association between an image and its respective textual description requires a specific study, which is out of the focus of this work.

As previously stated, our offer aggregation algorithm tries to minimize the intersection area between all the bounding boxes for the offers in the processed page. The algorithm starts by selecting the bounding box of a random Price P_i and merges it with its closest Description D_i and Title T_i bounding boxes to form an offer hypothesis O_i. The same process is repeated for all the remaining Prices in the page to form a finite set of hypotheses $H_{P_i} = \{O_0, \ldots, O_n\}$. The sum $S_{H_{P_i}}$ of the intersection areas between the bounding boxes for the offers in H_{P_i} is then calculated as follows:

$$S_{H_{P_i}} = \sum_{j \neq k} O_j \cap O_k, \;\; j, k \in \{0, \ldots, n\} \qquad (1)$$

Figure 4: Examples of flyers manually tagged by experts. The relevant entities listed in Table 2 are highlighted as coloured rectangles (Title, Description and Price).

This whole process is repeated multiple times, each time changing the starting seed Price, until all the Prices in the flyer have been selected as initial seeds. The set of offer hypotheses having minimum intra-intersection area is then selected as the best one.

4. EXPERIMENTS

In the remainder of this section we present the experimental results obtained testing the proposed method on marketing flyers randomly collected from different retailers. Throughout our experimental activity we evaluate quantitatively the accuracy of the method both at identifying and classifying entities of interest within the processed flyers and at aggregating the detected entities into offers.

4.1 Dataset

In order to evaluate the proposed approach, a total number of 797 product offers have been gathered from 103 marketing flyers produced by 2 different retailers. The collected documents come from electronic domains and present different design styles. On the ArteLab website, is available a zip file containing a subset of the entire dataset. To get the whole dataset please contact the authors of this paper.

Each flyer has been manually labelled by a team of 4 experts using a specially designed GUI. As shown in Fig. 4, the experts were instructed to provide both the coordinates of all the product Titles, Descriptions and Prices in the pages, and the associations between those bounding boxes and the different offers within the pages. The information gathered from the different experts has been averaged to obtain the final ground-truth data used to evaluate the proposed

method.

4.2 Evaluation Metrics

We evaluate the accuracy of the method both at classifying/aggregating individual tokens and at aggregating the merged tokens into product offers.

Since our ground-truth data is composed of labelled bounding boxes manually drawn by experts over the different flyers, we measure the accuracy of the proposed approach by evaluating the intersection-over-union (IoU) [3] score between the bounding boxes detected by the proposed approach and the respective ground-truth information.

Each entity is evaluated independently from the others. Given a page with its ground-truth data for one of the entities from Table 1, and the aggregated predictions provided by the model for the same entity class; the evaluation process for the token classification and aggregation phases is carried out as follows: we compare the IoU score between each ground-truth bounding box and the predictions provided by the model; if one of the predicted bounding boxes achieves an IoU score greater than 0.5 with the ground-truth bounding box, the prediction is considered correct. For every ground-truth bounding box at most one predicted bounding box might be considered correct. Given the number of correct predictions, we compute the classic Precision, Recall and F-measure values.

Given a page with its ground-truth offer data, and the offer hypotheses generated as in Sec. 3.2, the evaluation process for the offer aggregation phase is carried out as follows: we compare the IoU score between each component of a ground-truth offer (Title, Description and Price) and the bounding boxes for the same component in the offer hypotheses; if every predicted component for a given offer hypothesis has an IoU that is greater than 0.5 with its respective ground-truth offer component, then the predicted offer is considered correct. For every ground-truth offer at most one hypothesis might be considered correct.

4.3 Results

Tokens used to train the CNN were 188973 and the number of tokens used to test the model was 49887 which represents about 20% of the entire dataset. For each token, we compute the classification in the following way: first we get the page crop, which usually has a rectangular shape, then we do a squarify operation and resize images into new ones of 256x256 pixels to match the requested input of the AlexNet model. After all these steps we invoke the CNN to get the response according to our classes.

The first experiment aims at measuring the goodness of the proposed CNN classifier. As listed in Table 1, tokens may belong to one of four possible classes: Title, Description, Price and Other.

With the last experiment we evaluate the phases described in Sec. 3.1 and 3.2: the aggregation of tokens, and the subsequent aggregation of merged tokens into product offers. We measure Precision, Recall and F-measure values achieved on test set, while varying the token aggregation threshold ϵ from $0.1 \cdot \text{token\_height}$ to $10 \cdot \text{token\_height}$. We report the best obtained results in Table 2; they have been obtained setting $\epsilon = (2 \cdot \text{token\_height})$.
Even though the previous token classification/aggregation phase has not high accuracy values, in the last step we obtain good results because offer aggregation tends to group data into higher level of granularity.

Table 1: Confusion matrix for the CNN classifier.

	Title	Descr.	Price	Other
Title	85.31%	44.26%	1.87%	6.89%
Descr.	0.88%	37.81%	0.94%	1.81%
Price	0.11%	1.93%	73.54%	4.07%
Other	13.70%	16.00%	23.65%	87.24%

Table 2: Evaluation of both phases: token classification/aggregation and offer aggregation

	Precision	Recall	F-measure
Title	0.417	0.838	0.557
Description	0.495	0.807	0.614
Price	0.799	0.443	0.570
Aggr. offers	0.925	0.541	0.683

5. CONCLUSION

An ad-hoc method for the automatic extraction of structured product offers from marketing flyers has been proposed. The presented approach relies on CNN classification for the first step and then uses ad-hoc created algorithm for aggregating tokens and offers. The method has been evaluated over a collection of randomly collected flyers, achieving satisfying results while also maintaining an excellent language and genre independence due to the limited use of classical textual features.

6. REFERENCES

[1] E. Apostolova and N. Tomuro. Combining visual and textual features for information extraction from online flyers. In *EMNLP*, pages 1924–1929, 2014.

[2] K. Chellapilla, S. Puri, and P. Simard. High performance convolutional neural networks for document processing. In *10th Int. Workshop on Frontiers in Handwriting Recognition*. Suvisoft, 2006.

[3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge. *Computer Vision*, 88(2):303–338, 2010.

[4] I. Gallo, A. Zamberletti, and L. Noce. Content extraction from marketing flyers. In *Computer Analysis of Images and Patterns*, pages 325–336. Springer, 2015.

[5] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[7] P. Y. Simard, D. Steinkraus, and J. C. Platt. Best practices for convolutional neural networks applied to visual document analysis. In *ICDAR*, volume 3, pages 958–962, 2003.