Progress Report – Document image classification for digital library collections

08/13/2019
Mike Pack

Background

Document image classification aims to classify a type of given document image into a certain category—email, letter, handwritten, etc.—based on its layout and visual structure. A successful document image classification can breakdown and categorize a large-scale digital document repository into a smaller subset, which is beneficial for maintenance, discoverability, etc.

The main challenge of document image classification arises from the fact that within each document type, there exists a wide range of visual variability, as shown in Figure 1. Another issue is that documents of different categories often have substantial visual similarities, as shown in Figure 2.

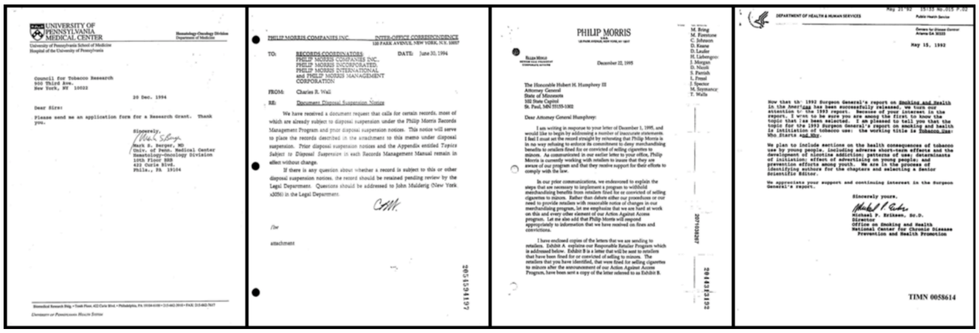

Figure 1. Examples of document images that show a wide range of visual variability within the same type (i.e., a *letter* type in this particular example). Note that no two documents show the exact same spatial arrangement of header, date, address, body, and signature; some of the documents even omit these components entirely.

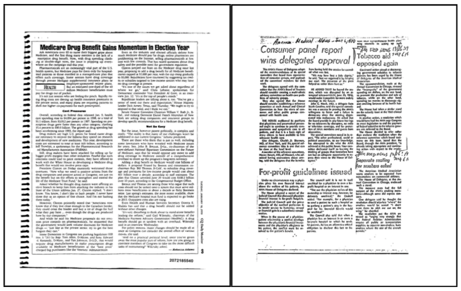

Figure 2. Examples of document images that show visual similarities across different types (Left: *form*, right: *scientific publication*) Note that two different types of documents share similar spatial arrangement of title and body. Even the amount of contents, so-called density, is also similar.

In the past few years, using a deep convolutional neural network (CNN) to classify images has shown to be able to achieves substantially successful classification performances in various domain, such as natural image classification, natural image segmentation, etc. Inspired by the success of CNNs in other domains, we would like to propose using the current state-of-the-art CNN model for document image classification problem.

State-of-the-Art

In this section, two papers, which used CNN for document image classification, are briefly reviewed. It is worth noting that all three papers used the same dataset, Ryerson Vision Lab Complex Document Information Processing (RVL_CDIP)[1] [1], and they achieved similar performance, around 90%.

Harley et al. (2015) investigated whether the features extracted from natural images (i.e., ImageNet) are general enough to be applied to document images [1]. The author also proposed a region-based CNN model, which consists of 5 different CNNs where each CNN is designed to be trained on particular regions: (1) holistic, (2) header, (3) footer, (4) left-body, and (4) right-body. Each of those CNN is VGG-16 [2] pre-trained on ImageNet. The dimension of each feature vector extracted from the corresponding CNN is reduced using principal component analysis, and they are concatenated into a single vector for the classification.

There are two interesting findings from their experimental result. First, *the features extracted from a CNN trained on ImageNet are powerful enough to be used for document image classification task that achieves approximately as well as a model fine-tuned on a subset of RVL_CDIP*, so-called *SmallTobacco*, 87.8% and 89.8%, respectively. There are two key implications from this finding. First, what the machine considers as distinctive features in natural images are also distinctive features in document images. Second, since we can easily transfer the knowledge (i.e., a set of filters capable of extracting distinctive features from an image) from one model to the other, we do not need to train our own model from scratch, which would allow us to reduce a significant amount of training time. Second, *given sufficient training data, enforcing region-specific feature-learning is unnecessary; a single CNN trained on entire images performed approximately as well as an ensemble of CNNs trained on specific subregions of document images*, 89.8% and 89.3%, respectively. This finding indicates that in the task of document classification, feeding large amount of data is more important than feeding fine-grained region-dependent representations. This result suggests that putting more efforts on collecting a larger amount of training dataset is *more* important than redesign a model's architecture for capturing a region-specific representation in the document classification.

Muhammad et al. (2017) investigated recent deep CNN architectures (i.e., AlexNet, VGG, GoogLeNet, and ResNet) and strategies for the task of document image classification [3]. Also, the author investigated the impact of transfer learning from a huge set of document images (i.e., RVL_CDIP). The outcome of this study can be summarized in two points as following: (1) VGG-16 performs slightly better than other networks by a small margin of 1-2%, and (2) with regards to the impact of transfer learning, all CNNs pre-trained on RVL_CDIP achieve higher accuracy than both ImageNet and random initialization (i.e., no transfer learning). The first outcome implies that there are no significant performance differences between recent CNN models, which allows one to use a computationally cost-effective model for practical deployment—if that is a concern. The second outcome is not a surprising result, which aligns with [1] in

[1] This dataset consists of 400,000 labeled document grayscale images from 16 classes. The images are sized, so their largest dimension does not exceed 1000 pixels.

that a model pre-trained on ImageNet outperforms a model trained from scratch. Overall, the key implication from this research is that using one of recent CNN model pre-trained on RVL_CDIP is a suitable preset for building our own document classification model.

Proposed Approach

As a first experiment for the task of document image classification, our goal is to build a model capable of distinguishing three different types of documents: (1) handwritten, (2) typed/machine-printed, and (3) mixed (both handwritten and typed). To this end, we propose to use a VGG-16 pre-trained on RVL_CDIP for the task of document image classification based on the two following findings: (1) a simple deep CNN architecture, especially VGG-16, showed better performance in the task of document classification than an ensemble model [1][3], and (2) a model pre-trained on RVL_CDIP outperformed both a model pre-trained on ImageNet and a model with random initialization.

The overall task can be detailed and broken down into two sub-tasks as below:

(1) Data acquisition: We first need to import datasets (i.e., campaigns) from By the People collection and manually label each image to construct a ground-truth. The number of data points in the smallest dataset in literature is 3,483 labeled images. So, hitting that number would be the best-case scenario. If this is not achievable, we can lower the bar to 1,000.

   a. **Subtask 1.** Write a script to download a bulk of images from LoC website using *loc.gov JSON API* to our *cdrhdev2* server.

   b. **Subtask 2.** Annotate each image with one of the following labels (integer format): (1) 0; handwritten, (2) 1; typed, and (3) 2; mixed.

(2) Training model: While we are doing the data acquisition, at the same time, we can setup and start Experiments 1 and 2. Once we have a dataset from the By the People collection, we can conduct Experiment 3.

   a. **Experiment 1.** In order to reproduce the results of aforementioned papers, we start training VGG-16 pre-trained on ImageNet with a *subset* of RVL_CDIP.

   b. **Experiment 2.** In order to generate a VGG-16 pre-trained on RVL_CDIP, we start training VGG-16 pre-trained on ImageNet with *full* RVL_CDIP.

   c. **Experiment 3.** We start training VGG-16 pre-trained on *full* RVL_CDIP with a dataset from By the People collection.

Reference

[1] Harley, A.W., Ufkes, A. and Derpanis, K.G., 2015, August. Evaluation of deep convolutional nets for document image classification and retrieval. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)* (pp. 991-995). IEEE.

[2] Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[3] Afzal, M.Z., Kölsch, A., Ahmed, S. and Liwicki, M., 2017, November. Cutting the error by half: Investigation of very deep cnn and advanced training strategies for document image classification. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (Vol. 1, pp. 883-888). IEEE.