Review of Evaluation of Deep Convolutional Nets for Document Image Classification and Retrieval

Dinesh Buswala

June 17, 2025

The paper presents the idea for automating document image and retrieval classification by leveraging the power of deep Convolution Neural Network (CNN). CNN is the current state of the art technique to learn the hierarchical chain of abstraction from pixel inputs to concise and descriptive representations.

There is much research work that happened in the past for document image analysis. There is a rigid structure in the document, which can be helpful in the document classification. However, this approach has tightly bound to the structure of the template and equivalent to a template matching problem. Also, this approach is limited to the documents which have a fixed design.

An alternative strategy is to treat document images holistically, or at least in vast regions, and search for discriminative landmark features that may appear anywhere in the document. This strategy is sometimes called a "bag of visual words". The advantage of holistic analysis is that the resulting representation of documents is invariant to the geometric configuration of the features.

There have been attempts to bridge the gap between region-based and holistic analyses. By concatenating image features pooled at several stages, beginning with a whole-image pool and proceeding into smaller and smaller regions, it is possible to build a descriptor that contains both global and local layout characteristics. This technique, known as spatial pyramid matching.

In structured documents, the layout of text and graphics elements often reflect essential information about the genre. The genre of the document is present in different regions of the document. For classifying or retrieval of the documents, we need to find these genres in the specified regions. Therefore, documents of a category often share region-specific features.

The paper follows two approaches to find the genre of the documents; the first one is to find the features from the document using single holistic CNN, and the second is to find the highlights from regions using the region-specific CNN and concatenating these features. There are few places in the documents where the information is rich like header, left body, right body, and footer.

The holistic CNN may not take advantage of the region-specific information, and this ap-

proach is more of the coarse grain approach. Whereas the region-specific CNN is taking advantage of the region-specific information, this approach is more of the fine-grain method. The key idea in the ensemble CNN is to learn the region-specific information independently.

A document can have many regions. However, few regions are very information-rich like header, left body, right body, and footer. Thus, in this work, a total of five CNNs are used. Four of these are region-tuned, placed at the header, left body, right body, and footer of the document images. The fifth is a holistic CNN, trained on the entire images.

All the images resized to 780 * 600 to generic the dimensions of the dataset. The header region describes in the first 256 rows of pixels in each image. The footer region describes in the last 256 rows of pixels in each image. The left body region was delineated by the intersection of the 400 central rows and the 300 remaining columns; the right body region was symmetrically defined. Every extracted region was resized to 227 * 227 before being used as input. Before building the network, the weights need to be initialized. There are two ways to initialize the weights first one is to initialize all the weights to zero, and the second one is to use the weights from the pre-trained network and fine-tune them while training.

A popular choice for pre-training is the ILSVRC 2012 ImageNet challenge. Features extracted from an ImageNet-trained network are compelling general-purpose features in a variety of other vision challenge. The result shows that the ensemble CNN out perform with all the available techniques.

The idea can be extended further to improve performance. Each region-specific CNN contains some information based on this information; the classification takes place. However, there are certain regions where the information is more important than other regions like headers have more information compared to other regions. The regions which carry more information are advantageous in the classification. We can leverage this fact by giving weights to each region based on the information the region has. By doing this, we can extend the control of overfitting and underfitting at one more level. Also, we can use some popular CNN architecture like VGG 16 and AlexNet to increase the performance.