2014 22nd International Conference on Pattern Recognition

Convolutional Neural Networks for Document Image Classification

Le Kang*, Jayant Kumar*, Peng Ye*, Yi Li†, David Doermann*
*University of Maryland, College Park, MD, USA
{lekang,jayant,pengye,doermann}@umiacs.umd.edu
†NICTA and ANU
yi.li@cecs.anu.edu.au

***Abstract*—This paper presents a Convolutional Neural Network (CNN) for document image classification. In particular, document image classes are defined by the structural similarity. Previous approaches rely on hand-crafted features for capturing structural information. In contrast, we propose to learn features from raw image pixels using CNN. The use of CNN is motivated by the the hierarchical nature of document layout. Equipped with rectified linear units and trained with dropout, our CNN performs well even when document layouts present large inner-class variations. Experiments on public challenging datasets demonstrate the effectiveness of the proposed approach.**

I. Introduction

Classifying and grouping document images into known categories is often a prerequisite step towards document understanding tasks, such as text recognition, document retrieval and information extraction [1]. These tasks can be greatly simplified if we know a priori the genre or the layout-type of documents. In the past, document image classification and retrieval has been done under a number of paradigms. Among which two major paradigms have been extensively studied: text-content based approaches and document structure based approaches. This paper follows the second paradigm and studies document structure based classification.

Previous approaches for document structure based classification have focused on finding effective visual representations. Existing approaches in the literatures differ from each other mainly in their choices of local features, global representations and learning mechanisms [2]. Various structure or layout-based features have been introduced [3], [4], [5], [6] and are shown to be effective for document image classification and retrieval. These approaches, however, are limited to a particular class of documents such as Bank forms, Memos, Contracts and Orders. In order to apply existing classification systems to other types of documents, we need to reconsider spatial features and tune it manually. Moreover, when the content and structure in documents are unconstrained as in handwritten documents, pre-defined features may not be able to capture all variations of a particular class.

A more general approach which automatically learns different abstractions of structure hierarchy and spatial relationship among document elements is desired. Document images usually have a hierarchical structure such as cells in rows and columns of tables, words in sentences, sentences in paragraphs. These hierarchical patterns are often repeated in different parts of document. These properties imply the possibility of learning the layout as a combination of small group of middle or lower level features.

In this paper, we present a general approach for document image classification using Convolutional Neural Networks (CNNs). CNN is a kind of neural networks that shares weights among neurons in the same layer. CNNs are good at discovering spatially local correlation by enforcing a local connectivity pattern between neurons of adjacent layers [7]. With multiple layers and pooling between layers, CNNs automatically learn the hierarchical layout features with tolerance to spatial translation, and by sharing weights it captures repeating patterns efficiently.

For the task of document image classification, a new type of neuron, Rectified Linear Units (ReLU) [8], is used in our CNN to speed up training. We employ dropout [9] to prevent overfitting. Experiments on real-world unconstrained datasets show that our approach is more effective than previous approaches.

II. Related Work

Byun and Lee [10] used a partial matching method in which document structure recognition and classification is applied to only part of input form images. The application of their approach is limited to *form* images and does not generalize to other types of documents. Shin and Doermann [11] defined *visual similarity* of layout structures and applied supervised classification for each specific type. They used image features such as the percentage of text and non-text (graphics, images, tables, and rulings) in content regions, column structures, relative point sizes of fonts, density of content area, and statistics of features of connected components. For classification, they used decision trees and self-organizing maps. Like previous approaches, the main drawback of their method is that the features were designed for specific document classes (e.g., forms, letters, articles). Additionally, due to a large number of different feature types the approach is computationally slow for large scale document exploration.

Collins-Thompson and Nickolov [12] proposed a model for estimating the inter-page similarity in ordered collections

The partial support of this research by DARPA through BBN/DARPA Award HR0011-08-C-0004 under subcontract 9500009235, the US Government through NSF Awards IIS-0812111 and IIS-1262122 is gratefully acknowledged.

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.546

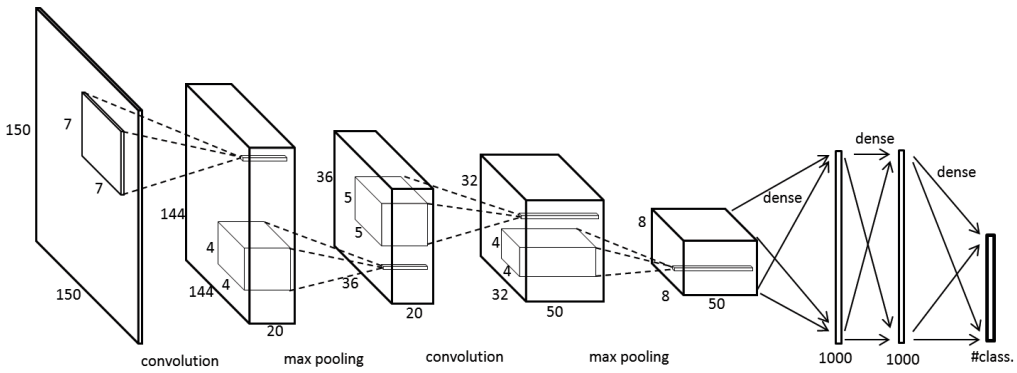


Fig. 1: The architecture of the proposed CNN

of document images. They used features based on a combination of text and layout features, document structure, and topic concepts to discriminate between related and unrelated pages. Since the text from OCR may contain errors, especially for handwritten documents, the approach is limited to well-structured printed documents. Joutel et al. [13] presented an approach for the retrieval of handwritten historical documents at page level based on the *curvelet transform* to compose a unique signature for each page. The approach is effective when local shapes are important for classification but the approach is likely to miss any higher level of structural saliency. In many cases, the desired similarity between document images is embedded in page layout or structure.

Kochi and Saitoh [14] proposed a system for identifying the type of a semi-formatted document based on important textual elements extraction and by using a flexible matching strategy for easy model generation. Bagdanov and Worring [3] approached the general problem of genre classification of printed document images using attributed relational graphs (ARGs). They used ARGs to represent the layout structure of document instances, and the first order random graphs (FORGs) to represent document genres. They reported a high-accuracy on a small dataset of 130 documents consisting of 10 genres. Reddy et al. [15] addressed the form classification problem with a classifier based on the k-means algorithm. They used low-level pixel density features and adaptive boosting to classify NIST tax forms. A detailed survey on document classification based on three components: the problem statement, the classifier architecture, and the performance evaluation can be found in Chen and Blostein [2].

Approaches based on bag-of-words (BOW) models have shown promising results on many computer vision problems, such as image classification [16], scene understanding [17] and document image classification [18], [19]. However, initial formulations typically disregard the spatial relationships different image regions, and only consider the occurrences of visual patterns in an image. This results in a limited descriptive capability and the performance may drop significantly in presence of noise, background clutter, variation of layout and content in images. Subsequently, methods which extend the BOW approach to incorporate spatial relationships between image regions have been proposed. One of the early methods proposes the creation of spatial-pyramid features by partitioning the image into increasingly finer grids and computing the weighted histogram based kernel in each region [20]. Recently, there has been a focus on selecting the optimal feature combination strategy and efficient ways to learn these local statistics, and a number of methods have been proposed [21], [22]. Kumar et al. [4], [6] extended the spatial-pyramid features for document images by using a novel pooling method with horizontal-vertical partitions that are adapted to the typical layout of document images.

III. CNN for Document Image Classification

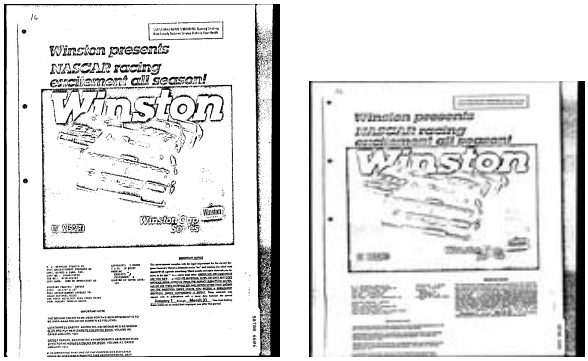
We propose to use a CNN for document image classification. The main idea is to learn a hierarchy of feature detectors and train a nonlinear classifier to identify complex document layouts. Given a document image, we first perform downsampling and pixel value normalization, then feed the normalized image to the CNN to predict the class label.

A. Preprocessing

The resolution of document images is typically higher than 2000×2000, which is too large to be fed to a CNN with current availability of computing resources. Large input dimension not only costs more computation resources but also leads to greater chance of overfitting. Considering the fact that it is the layout, instead of the details such as characters, that determines the class of document images, we can reduce the input dimension by discarding details of document images as long as the structure information is still identifiable. Specifically, document images of various sizes are all downsampled and resized to 150×150 with bilinear interpolation. At the resolution of 150×150, most characters on the document images are not recognizable but the overall layout is preserved and the locations of title, text or table can be determined. Humans can still make predictions on the document types no worse than at original resolution if judging by layout only. Fig. 2 shows the downsampled document images compared to original resolution. After downsampling, the gray scale images are divided by 255 and then subtracted by 0.5, therefor normalized to the range of $[-0.5, 0.5]$.

B. Network Architecture

Fig. 1 shows the architecture of our network, which can be summarized as $150 \times 150 - 36 \times 36 \times 20 - 8 \times 8 \times 50 - 1000 -$



(a) (b)

Fig. 2: (a) Original image of resolution 2544×3256 (b) Downsampled and resized to 150×150.

$1000 - M$, where M is the number of classes. The input is downsampled and normalized image of size 150×150 . The first convolutional layer consists of 20 kernels, each of size 7×7, followed by a 4×4 pooling that reduces the each feature map to a 36×36. The second convolutional layer contains 50 kernels each of size 5 which means each kernel is convolved with all 20 feature maps of previous layer. A 4×4 pooling comes after the second convolutional layer to produce 50 feature maps each of size 8×8. Two fully connected layers of 1000 nodes each follow the convolution and pooling layers. The last layer is a logistic regression with softmax that outputs the probability on each class, as defined in the following equation

$$P(y = i|x, W_1, ..., W_M, b_1, ..., b_M) = \frac{e^{W_i x + b_i}}{\sum_{j=1}^{M} e^{W_j x + b_j}} \quad (1)$$

where x is the output of the second fully connected layer, W_i and b_i are the weights and biases of i^{th} neuron in this layer, and M is the number of classes. The class that outputs the max probability is taken as the predicted class, which can be described in the following equation ($\hat{y}$ denotes the predicted class)

$$\hat{y} = \arg\max_i P(y = i|x, W_1, ..., W_M, b_1, ..., b_M) \quad (2)$$

Instead of traditional sigmoid or tanh neurons, we use Rectified Linear Units (ReLUs) [8] in the convolutional and fully connected layers. Recently research [23] demonstrated ReLUs brings several times speedup in training compared to using tanh units. Formally, an ReLU has an output of $f(x) = max(0, x)$ where x denotes the input. In experiments we observe that ReLUs enable the training to complete several times faster and not so sensitive to the scale of input.

C. Training

We adopt negative log-likelihood as the loss function and perform Stochastic Gradient Descent (SGD). Recently successful neural network methods report that dropout [23], [9] improves learning. During training time the neuron outputs are masked out with probability of 0.5, and at test time their outputs are halved. Dropout alleviates overfitting by introducing random noise to training samples. In our experiment we also find dropout boosts the performance for a large network. Since applying dropout to all layers significantly increases the training time to reach convergence, we only apply dropout at the second fully connected layer, i.e., half of the outputs of the second fully connected layer are randomly masked out in training, and in testing the weights of the logistic regression layer are divided by 2, which is equivalent to halving the outputs of the second fully connected layer.

IV. Experiment

We conduct experiments on two datasets to demonstrate the effectiveness of our CNN.

A. Datasets

The following two datasets were used in our experiments.

(1) Tobacco litigation dataset [24]: we used 3482 images categorized in 10 genres(classes): *report, memo, resume, scientific, letter, news, note, ad, form, email*. Fig. 3 shows some samples of Tabacco dataset. From Fig. 3 we can see that there is large inner-class variation, especially for the class *ad*.

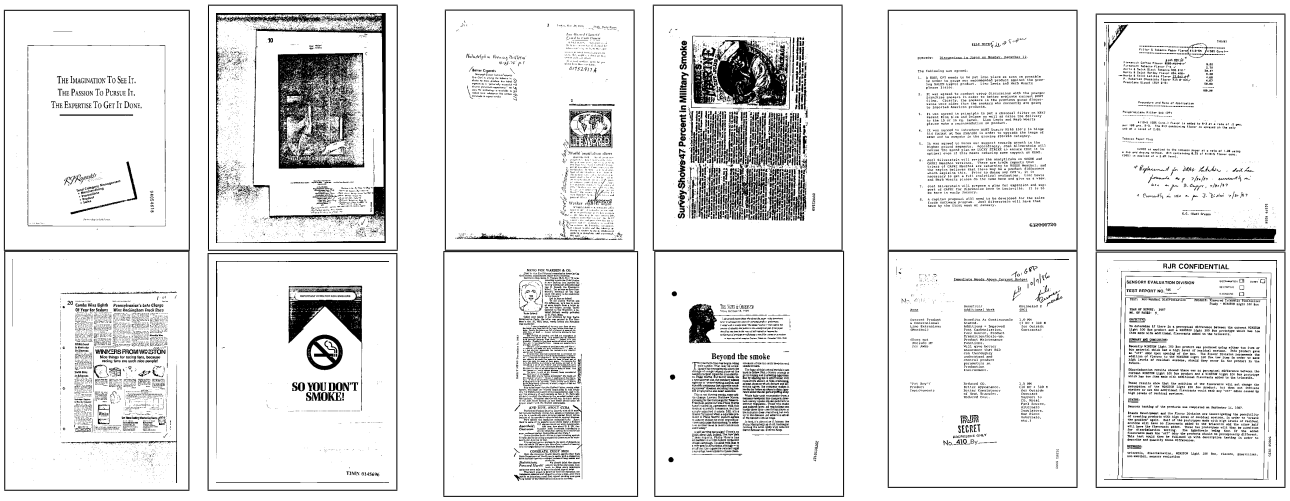
(2) NIST tax-form dataset [25]: a collection of 5590 tax-form images from National Institute of Standards and Technology, categorized into 20 classes, with labels like *Form1040-1, Form1040-2, Form4562-1, Form2441* and so on. Fig. 4 shows samples of NIST tax-forms.

B. Evaluation

We mainly compare our method to previous methods spatial pyramid/random forest (SP-RF) and horizontal-vertical partitioning/random forest (HVP-RF) [6], therefore we follow the same evaluation protocol. We apply the proposed CNN with the same architecture to the two datasets described above.

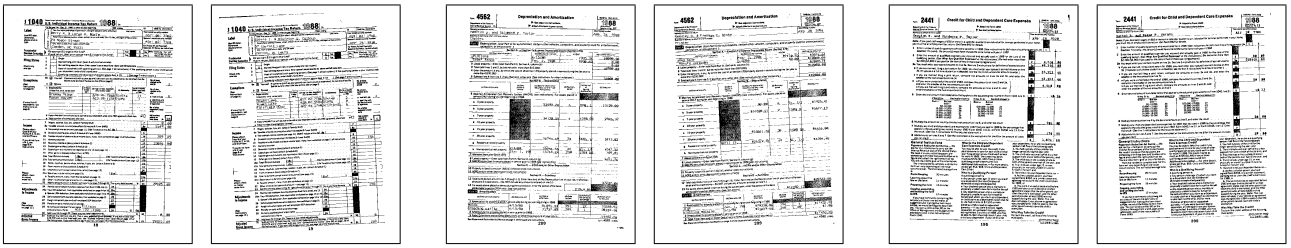
For 10 classes of images in the Tobacco dataset, we randomly select $N (N <= 100)$ images per class for training and validation, among which 80% are for training and 20% for validation, and the rest images are used for test. We vary N to see the performance under different amount of training and validation samples. The accuracies of the proposed algorithm are obtained on 100 such random partitions of training, validation and test, and the median accuracy is shown in Fig. 5. The proposed approach achieves a median accuracy of 65.37% when 100 samples are used for training and validation. Our CNN consistently outperforms SP-RF and HVP-RF [6]. A class-confusion matrix on one of the partitions is shown in Table I.

On the 20-class NIST tax-form dataset, we randomly pick one image per class (which amounts to 20 samples in total) for training, and use the rest for test. Validation set is not used here. We simply use the parameters after 50 epochs of training. A median accuracy of of 100% is achieved through 100 partitions of training and test, which ties with [6]. Other methods such as [26] achieved similar accuracies, but more training samples are used. We believe that the proposed method



(a) *ad* (b) *news* (c) *report*

Fig. 3: Sample images from Tobacco dataset, grouped in three genres/classes (a) *ad*, (b) *news* and (c) *report*



(a) *Form1040-1* (b) *Form4562-1* (c) *Form2441*

Fig. 4: Sample images from from NIST tax-form dataset, grouped in three classes (a) *Form1040-1*, (b) *Form4562-1*, and (c) *Form2441*

	ad	email	form	letter	memo	news	note	report	resume	scientific
ad	**104**	0	1	1	0	9	2	2	0	3
email	1	**435**	7	3	13	0	4	3	1	0
form	2	0	**145**	5	37	7	8	7	0	14
letter	0	8	6	**297**	43	0	1	14	0	10
memo	1	7	33	51	**294**	6	3	9	0	18
news	19	1	21	13	6	**45**	8	2	0	16
note	2	10	24	8	31	5	**63**	0	0	11
report	1	15	34	65	32	11	5	**103**	5	38
resume	0	7	24	13	12	1	1	13	**13**	6
scientific	0	16	36	11	52	4	6	12	1	**45**
Accuracy (%)	80.0	87.2	43.8	63.6	56.6	51.1	62.4	62.4	65.0	28.0

TABLE I: Class-confusion matrix for genre classification on Tobacco dataset. This is the results of one partition of training-validation-test, which gives an overall accuracy of 65.35%

achieves such high accuracies with so few training samples because the tax-form images in the same class show highly consistent layout and inter class similarity is relatively low.

We visualize the kernels of the first convolutional layer learned on the Tobacco and the NIST respectively, as shown in Fig. 6. We do not observe obvious patterns that resemble the local structure of document images. But this is not surprising since our approach is purely supervised and does not aim to learn visually appealing features that generative models typically use.

C. Computational Cost

We implemented the CNN using the python library Theano [27] which enables easy deployment on a GPU to speed up

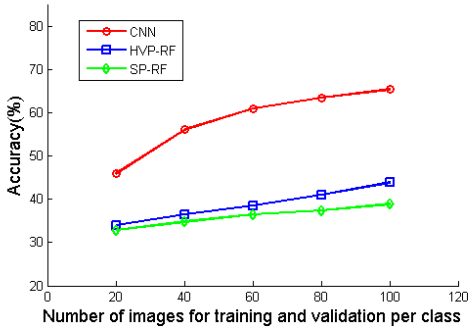
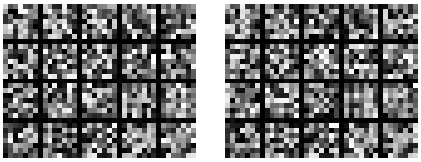


Fig. 5: Classification results on Tabacco dataset



(a) (b)

Fig. 6: Learned kernels in the first convolutional layer from (a) Tobacco dataset (b) NIST dataset.

the process without much manual optimization. Our experiments were performed on a PC with 2.8GHz CPU and Tesla C1060 GPU. On average each image takes about 0.004 second processing time.

V. Conclusion

We have proposed a convolutional neural network for document image classification. Contrary to traditional methods that use hand-crafted features, our algorithm learns hierarchical features directly from normalized image pixels. We employ ReLU and dropout to enhance the training of our CNN. Experiments on public datasets show that our algorithm achieved state of the art performance.

References

[1] A. Dengel and F. Dubiel, "Clustering and classification of document structure-a machine learning approach," in *International Conference on Document Analysis and Recognition*, vol. 2, 1995, pp. 587–591.

[2] N. Chen and D. Blostein, "A survey of document image classification: problem statement, classifier architecture and performance evaluation," *International Journal Document Analysis Recognition*, vol. 10, no. 1, pp. 1–16, 2007.

[3] A. D. Bagdanov and M. Worring, "Fine-grained document genre classification using first order random graphs," in *International Conference on Document Analysis and Recognition*. IEEE, 2001, pp. 79–83.

[4] J. Kumar, P. Ye, and D. Doermann, "Learning Document Structure for Retrieval and Classification," in *International Conference on Pattern Recognition*, 2012, pp. 1558–1561.

[5] S. Chen, Y. He, J. Sun, and S. Naoi, "Structured document classification by matching local salient features," in *International Conference on Pattern Recognition*, 2012, pp. 653–656.

[6] J. Kumar, P. Ye, and D. Doermann, "Structural similarity for document image classification and retrieval," *Pattern Recognition Letters*, 2013.

[7] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *IEEE proceedings*, 1998, pp. 2278–2324.

[8] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *International Conference on Machine Learning*, 2010, pp. 807–814.

[9] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv:1207.0580*, 2012.

[10] Y. Byun and Y. Lee, "Form classification using dp matching," in *Proceedings of the 2000 ACM Symposium on Applied Computing - Volume 1*, 2000, pp. 1–4.

[11] C. Shin and D. Doermann, "Document image retrieval based on layout structural similarity," in *International Conference on Image Processing, Computer Vision and Pattern Recognition*, 2006, pp. 606 – 612.

[12] K. Collins-Thompson and R. Nickolov, "A clustering-based algorithm for automatic document separation," in *SIGIR Workshop on Information Retrieval and OCR: From Converting Content to Grasping, Meaning*, 2002, pp. 1–8.

[13] G. Joutel, V. Eglin, S. Bres, and H. Emptoz, "Curvelets based queries for cbir application in handwriting collections," in *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, vol. 2, 2007, pp. 649–653.

[14] T. Kochi and T. Saitoh, "User-defined template for identifying document type and extracting information from documents," in *International Conference on Document Analysis and Recognition*, 1999, pp. 127–130.

[15] K. V. U. Reddy and V. Govindaraju, "Form classification," in *SPIE Document recognition and Retreival*, vol. 6815, 2008, pp. 1–6.

[16] C. Wallraven, B. Caputo, and A. Graf, "Recognition with local features: the kernel recipe," in *IEEE International Conference on Computer Vision*, 2003, pp. 257–264.

[17] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. Van Gool, "Modeling scenes with local descriptors and latent aspects," in *International Conference on Computer Vision*, vol. 1, 2005, pp. 883–890.

[18] E. Barbu, P. Héroux, S. Adam, and É. Trupin, "Using bags of symbols for automatic indexing of graphical document image databases," *Ten Years Review and Future Perspectives: Graphics Recognition*, pp. 195–205, 2006.

[19] J. Kumar, R. Prasad, H. Cao, W. Abd-Almageed, D. Doermann, and P. Natarajan, "Shape Codebook based Handwritten and Machine Printed Text Zone Extraction," in *International Conference on Document Recognition and Retrieval*, 2011, pp. 7874:1–8.

[20] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 2169 – 2178.

[21] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1794 –1801.

[22] Y. Yang and S. Newsam, "Spatial pyramid co-occurrence for image classification," in *International Conference on Computer Vision*, 2011, pp. 1465 –1472.

[23] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[24] D. Lewis, G. Agam, S. Argamon, O. Frieder, D. Grossman, and J. Heard, "Building a test collection for complex document information processing," vol. 2006, 2006, pp. 665–666.

[25] D. Dimmick, M. Garris, and C. Wilson, "Nist structured forms reference set of binary images (sfrs)," 1991. [Online]. Available: ⟨http://www.nist.gov/srd/nistsd2.cfm⟩

[26] S. Chen, Y. He, J. Sun, and S. Naoi, "Structured document classification by matching local salient features," in *International Conference on Pattern Recognition*. IEEE, 2012, pp. 653–656.

[27] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, "Theano: a CPU and GPU math expression compiler," in *Proceedings of the Python for Scientific Computing Conference (SciPy)*, Jun. 2010.