IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 17, 2019, accepted August 31, 2019, date of publication September 18, 2019,
date of current version September 27, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2940884

Unsupervised Exemplar-Based Learning for Improved Document Image Classification

SHERIF ABUELWAFA, MARCO PEDERSOLI, AND MOHAMED CHERIET, (Senior Member, IEEE)
École de technologie supérieure, University of Quebec, Montreal, QC H3C 1K3, Canada
Corresponding author: Sherif Abuelwafa (sherif.abuelwafa.1@ens.etsmtl.ca)

This work was supported by the NSERC of Canada under Grant RGPIN 2014-04649 and Grant RGPIN 2018-04825.

ABSTRACT Many recent state-of-the-art approaches for document image classification are based on supervised feature learning that requires a large amount of labeled training data. In real-world problem of document image classification, the available amount of labeled data is limited and scarce while a large amount of unlabeled data is often available at almost no cost. In this paper, we present an approach for learning visual features for document analysis in an unsupervised way, which improves the document image classification performance without increasing the amount of annotated data. The proposed approach trains a neural network model on an auxiliary task in which every training example is associated with a different label (exemplar) and expanded to multiple images through a data augmentation technique. Thus, the learned model, which is trained in an unsupervised way, is used to boost the document classification performance. In fact, this learned model has proved to be consistently efficient in two different settings: i) as an unsupervised feature extractor to represent document images for an unsupervised classification task (i.e., clustering); and ii) in the parameters initialization of a supervised classification task trained with a small amount of annotated data. We perform experiments on the Tobacco-3482 dataset and demonstrate the capability of our approach to improve i) the unsupervised classification accuracy up to 2.4%; and ii) the supervised classification accuracy by 1.5% without any extra data or by 5% when using 3000 additional not annotated samples.

INDEX TERMS Document image classification, document analysis, document image representation.

I. INTRODUCTION

Document image classification is a crucial step in the process of document understanding. Finding the document category is essential to later understanding steps, such as text recognition and document retrieval [1]. The current state-of-the-art approaches for document image classification depend on either carefully hand-crafted features [2]–[4] or feature learning [5]–[9]. Engineering features is a complex process that requires special expertise for designing and adapting the features to the desired domain and makes it hard to generalize to new tasks [10], [11]. Recently, approaches that directly learn features from data have received more interest and it is also the approach that we use. Among the feature learning approaches, methods based on Convolutional Neural Networks (CNNs), in which features are learned by the convolutional layers [5]–[9], achieved state of the art performance.

In terms of supervision, most of the successful feature learning approaches in the domain of document image classification are based on a supervised pre-training paradigm. Using fully supervised feature learning is often an efficient solution that provides very good results as long as enough labeled training data can be provided. This is not often the case in document classification, because the process of manually annotating data is slow and expensive in terms of both, the needed time and expertise. This results in a limited amount of labeled data that can actually be used in the feature learning process. On the other hand, a large amount of related unlabeled data is widely available (e.g., HathiTrust digital library[1] that contains millions of digitized document images).

Thus, semi-supervised and unsupervised approaches seem to be a good solution to improve the classification results without increasing the amount of annotated data. For instance, unsupervised feature learning approaches at the pre-training stage [12], [13] can provide substantial

The associate editor coordinating the review of this manuscript and approving it for publication was Donato Impedovo.

[1] https://www.hathitrust.org/

classification improvements [14] without additional data annotation. In these approaches, structural and spatial-related features are learned using only unlabeled data. Then, these learned features are used at a later fine-tuning stage, improving the supervised classification performance.

In this paper, we propose to first learn a neural network model during a pre-training phase on a set of data without annotation, thus in an unsupervised manner. This is performed with an exemplar learning in which a neural network is trained to accomplish the auxiliary task of classifying each sample in a data-augmented version of the original dataset. Then, we tackle the problem of document image classification, in which the pre-trained model is used in two different ways: i) in an unsupervised manner, by clustering on features extracted with the pre-trained model ii) initializing a supervised training with the pre-trained network weights. In both cases, the pre-trained model consistently improves the classification performance, over the baseline approaches on the respective tasks, without the need to use any additional labeled data. Note that for unsupervised classification, the reported results are with respect to a baseline that does not utilize the learned featured of our pre-trained network; in addition to, other methods based on a more complex clustering algorithm and hand-crafted features. For the supervised case, the baseline to compare with is a trained model without our pre-trained initialization of the weights.

A. CONTRIBUTIONS OF THIS PAPER

Our paper provides the following contributions:

- We propose a unified unsupervised pre-training based framework that is simple, yet capable of consistently boosting the performance of both unsupervised and supervised classification.
- To the best of our knowledge, our approach is the first to perform an unsupervised document image classification using a representation that is entirely based on feature learning using unlabeled data, and does not depend on any hand-crafted features. In the experimental results, we show that our approach outperforms the previous baseline approaches.
- We demonstrate and experimentally validate that by incorporating a small fraction of unlabeled data from a related-dataset we can easily gain up to 2.4% boost in the unsupervised classification performance and over 5% boost in the supervised classification performance.

The organization of this paper is as follows; section II provides a comprehensive review study on the related work. In section III, the proposed approach is introduced in details. The experimental setup is presented at section IV and the results with their related analysis are discussed at section V. Finally, the paper is concluded at section VI.

II. RELATED WORK

A. DOCUMENT IMAGE CLASSIFICATION

The problem of document image classification has been tackled in the literature through many approaches that differentiate based on i) the chosen features and ii) the utilized learning mechanism [15].

Considering the chosen features, recent approaches in the literature are either content (text) based [16], visual appearance based or a combination of both [17]. The content-based approaches are typically restricted to documents with text and depend mainly on Optical Character Recognition (OCR) methods, which may output text with errors that can affect the classification performance [18]. To avoid this, our proposed work is based instead on the visual appearance characteristics of the document image and does not rely on OCR.

Conventionally, visual appearance based approaches utilizes hand-crafted features [2]. For instance, Scale Invariant Feature Transform (SIFT) [19] is exploited by [3] and Speeded Up Robust Features (SURF) [20] is used by [21]. However, lately, visual appearance approaches that are based on feature learning [7] have attracted considerable attention.

Since this work is mainly focused on the pre-training stage of the classification process, a review on the related visual appearance based works is discussed in further detail below.

The simplest and most used pre-training approach that has been utilized extensively in recent years is supervised pre-training, in which a big and fully labeled dataset is used to perform a pre-training process [10]. For instance, [5], [6], [8], [9] are all incorporating a supervised pre-training process. In this process, annotated samples are used to train a network in a supervised manner, then that pre-trained network's learned parameters are used to initialize a fine-tuning network and perform the process of document classification. Usually a huge amount of annotated data is exploited in this process; for example, around 1 million labeled images of ImageNet [22] are used in [6], [8], and [9] and 320,000 labeled images of RVL-CDIP [6] are used in [5]. Das *et al.* [8] extended that approach using an ensemble of region-based classifiers (i.e., a strategy that has been introduced by [23]). However, the method is still limited to a specific set of documents (e.g., forms, memo), and cannot be applied easily to other document types because it depends on the spatial features of the documents and requires a manual readjustment to the learning algorithm for any new document type. Similarly, [6] evaluated enforcing learning region-specific features and concluded that it is not effective in case of enough training data.

In addition to supervised classification, some related works in literature have explored classifying the document images in an unsupervised manner, hence considered as 'unsupervised classification' (i.e., more details about the unsupervised classification process are discussed at section III-C). For instance, [4] introduced the horizontal vertical partitioning-random forest (HVP-RF) model, which trains a random forest classifier to learn structural patterns from SURF features [20] codebook. This model has a complex pipeline that depends heavily on traditional hand-crafted features; in contrast, our approach achieves better results using a pipeline that is based entirely on unsupervised feature learning. Moreover, the CONFIRM algorithm [24] uses page elements such as OCR transcriptions and rule lines to obtain collection-dependent features.

Using rule lines makes this approach limited and more specific to tables. Additionally, depending on OCR is not ideal as discussed earlier in this section.

B. UNSUPERVISED FEATURE LEARNING

Unsupervised feature learning often works on modeling the distribution of the training data to learn the common invariant features in it. For instance, Deep Belief Networks (DBNs) [25] learn features by yielding the parameters that maximize the latent variables likelihood given the observed ones. The main drawback of this technique is its inefficiency due to the intractability of the estimation of the latent variables likelihood. On the other hand, in direct mapping techniques, features are learned by minimizing the error between an input sample and the reconstructed output or some variants of it (e.g., stacked denoising auto-encoders [26], k-sparse auto-encoder [27] and variational auto-encoder [28]). Another interesting approach for improving the classification accuracy of documents is to perform an unsupervised pre-training. On the contrary to the supervised approach, the unsupervised pre-training depends only on unlabeled data. This means fast and cheap access to the available data since the labeling process has been bypassed. Even if very appealing, the impact of unsupervised pre-training on the final classification performance is still limited and not performing as effective as supervised pre-training.

A special case of unsupervised pre-training, is *self-supervised pre-training*. In that case, the learning task exploits the structure of the training data, such that data annotations are already available or come for free. In this way, normal supervised learning techniques can be used on those pseudo-annotations. For instance, the spatial information of neighboring patches is used to automatically label the input data through either context prediction [29] or solving jigsaw puzzles [30]. Additionally, [31] applies four different rotations to each unlabeled sample and trains a network to recognize the correct one. On the same line, an exemplar-based learning with CNN is introduced by Dosovitskiy *et al.* [32]. In this approach, data-augmentation is applied to each unlabeled sample to create a set of surrogate classes and a network is trained to discriminate between them. Due to its simplicity and closeness to the classification tasks, the Exemplar-CNN based learning has inspired the pre-training part of our framework. However, various changes in the architecture have been introduced for better adaptation to the problem of structural document classification.

III. THE PROPOSED METHODOLOGY

Our proposed framework is based on an unsupervised pre-training step in which a convolutional neural network (CNN) model is learned using only unlabeled data. This is followed by two different document image classification approaches: an unsupervised classification on the learned representation and a supervised classification initialized with the pre-trained model.

TABLE 1. The architecture of the CNN model used in our experiments.

Layer (type)	Output shape	Filter
input (InputLayer)	1 x 227 x 227	-
conv_1 (Conv2D)	96 x 55 x 55	11 x 11
max_pooling_1 (MaxPooling2D)	96 x 27 x 27	3 x 3
conv_2 (Conv2D)	256 x 27 x 27	5 x 5
max_pooling_2 (MaxPooling2D)	256 x 13 x 13	3 x 3
conv_3 (Conv2D)	384 x 13 x 13	3 x 3
conv_4 (Conv2D)	384 x 13 x 13	3 x 3
conv_5 (Conv2D)	256 x 13 x 13	3 x 3
max_pooling_3 (MaxPooling2D)	256 x 6 x 6	3 x 3
flatten (Flatten)	9216	-
dense_1 (Fully-connected)	4096	-
dense_2 (Fully-connected)	4096	-
dense_3 (Fully-connected)	N*	-

* Number of surrogate classes.

More insights on the different learning stages and other related steps are detailed in the following subsections.

A. PRE-PROCESSING

As shown in table 1, our network is based on the AlexNet architecture [33]. Thus, in order to match the network input size, all the utilized input document images, at the stages of unsupervised pre-training and classification, are resized to 227x227 pixel resolution. To provide an efficient processing performance, the resizing process keeps the fundamental structural features of the document, while reducing other less critical information for our model (e.g., the exact shape of characters and words). After resizing the image, a binarization process is performed: the image pixels values are rounded to either 0 or 1.

B. UNSUPERVISED PRE-TRAINING STAGE

The main objective of this stage is to train a CNN model using a set of unlabeled data. As shown in Fig. 2, the training procedure is composed of two steps: first, the generation of augmented data and surrogate classes; and then the actual training of the neural network to classify these generated classes. The two steps are detailed in the following paragraphs.

1) GENERATE AUGMENTED DATA AND SURROGATE CLASSES

Inspired by data augmentation [34] and similarly to [32], we generate a set of transformations of our original document images such that the augmented data are still valid and realistic document representations. We consider an initial training set $\mathcal{X}$ containing N unlabeled document images. A set of randomly-chosen combination of pre-defined transformations $\{T_1, \ldots T_K\}$ is applied to each image $x_i \in \mathcal{X}$, which produces K augmented versions of this image. Specifically, each augmented image $T_k x_i$ is the result of incrementally applying (with 50% probabilities) three basic transformations. To guarantee robust, descriptive and generic learned features, the following basic transformations that relate to some core characteristics of the document images have been

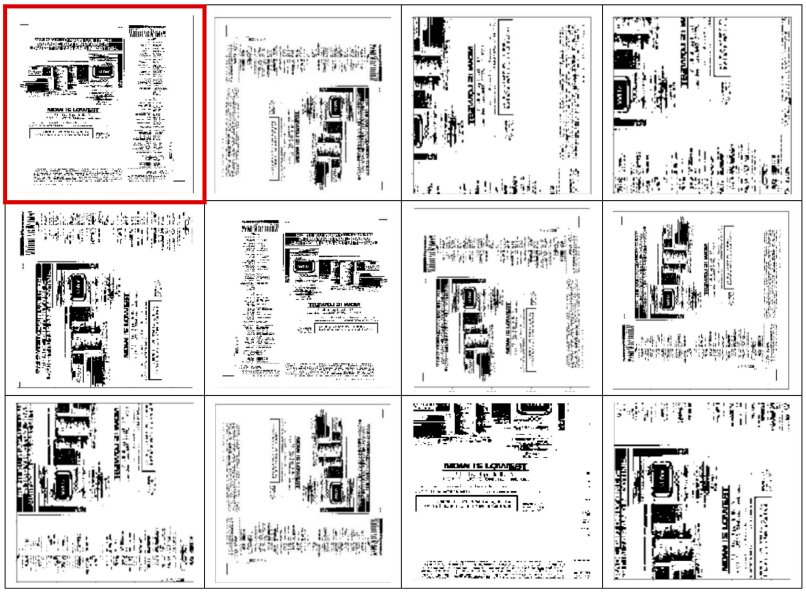


FIGURE 1. Applying T transformations (i.e., rotation by angles $\pm 90°$, zooming-in by a uniformly sampled factor between 1 and 1.15, and horizontal flipping) to an unlabeled document image x_i from the Tobacco-3482 dataset to generate $\{T_1 x_i, \ldots T_K x_i\}$ samples of a surrogate class S_{x_i}. The seed image x_i is at the top left corner.

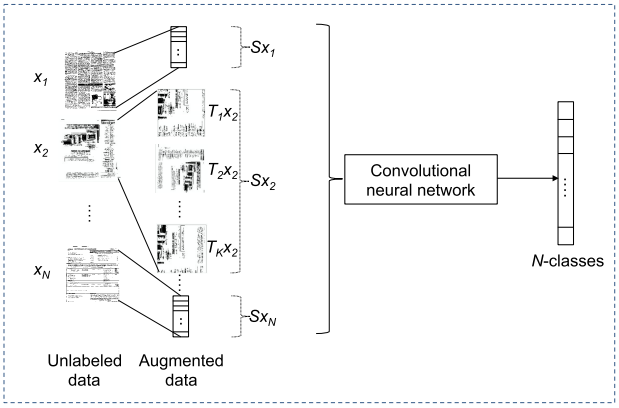


FIGURE 2. Proposed unsupervised pre-training stage.

used: rotation by angles 90 or -90 degrees, zooming-in by a uniformly sampled factor between 1 and 1.15, and horizontal flipping. Algorithm 1 provides more details on how the augmentation process is carried out. Each unlabeled image, x_i, is now considered a surrogate class S_{x_i}, and its corresponding generated transformations $\{T_1 x_i, \ldots T_K x_i\}$ are samples of that class with a surrogate label $i \in N$. Fig. 1 shows some generated samples of a surrogate class. We will show that the numbers of surrogate classes N and samples per surrogate class K have a critical impact on the classification performance; more insights are discussed in subsection V-B1.

2) TRAIN THE NETWORK

An exemplar learning process is accomplished using the obtained set of N surrogate classes and their $N * K$ samples. Specifically, a neural network is trained to associate each sample $T_k x_i$ to its related surrogate class S_{x_i} by minimizing the augmented samples cross-entropy loss:

$$L(X) = \sum_{i=1}^{N} \sum_{k=1}^{K} l(T_k x_i, i),$$
$$l(x, i) = -log(p(y = i; x)), \quad (1)$$

Algorithm 1 Generate Surrogate Classes: for Each Image x_i, we Generate a Transformation as a Random Composition of Rotation of θ Degrees (R_θ), Zoom-in by a Factor z (Z_z) and Horizontal Flip (F)

for each $x_i \in \mathcal{X}$ **do**
  for $k = 1$ to K **do**
    $T_k = I$ ▷ I: identity transf.
    rotate $\sim$ *Bernoulli*(0.5)
    if *rotate* **then**
      $\theta \leftarrow$ either $-90°$ or $90°$
      $T_k = T_k \circ R_\theta$
    end if
    zoom-in $\sim$ *Bernoulli*(0.5)
    if *zoom-in* **then**
      $z \sim U(1, 1.15)$
      $T_k = T_k \circ Z_z$
    end if
    flip $\sim$ *Bernoulli*(0.5)
    if *flip* **then**
      $T_k = T_k \circ F$
    end if
  end for
  $S_{x_i} = \{T_1 x_i, \ldots, T_K x_i\}$
end for

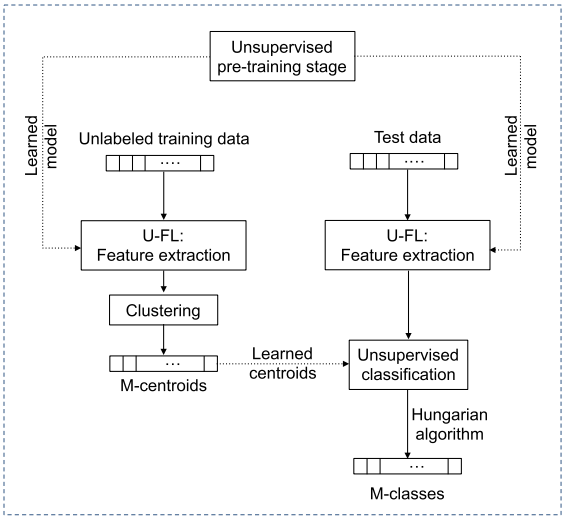


FIGURE 3. Proposed unsupervised classification stage.

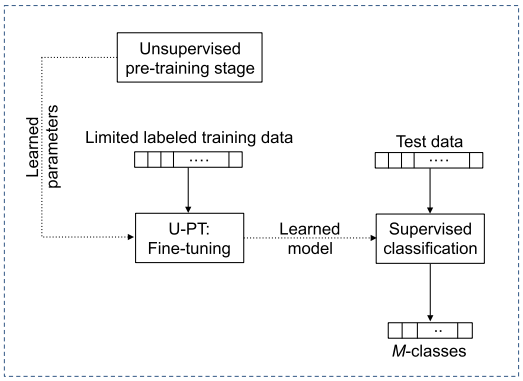


FIGURE 4. Proposed supervised classification stage.

where $p(y = i; x)$ is the probability of sample x to belong to class i and $p(.)$ is the softmax output of our network. After training, the obtained network parameters θ are considered to be invariant to the transformations used during the augmentation process.

The used network, as reported in table 1, contains eight layers (i.e., five convolutional and three fully connected layers) with around 56 million parameters. A zero padding is included to all the convolutional layers except the last one. In addition, the last fully-connected layer is coupled with an N-way softmax that provides an estimate of each class's conditional probability.

C. UNSUPERVISED CLASSIFICATION STAGE

As illustrated in Fig. 3, the unsupervised classification is actually a clustering process in its core. During training, we divide the training data into clusters and then associate each cluster to the best class in the test data. Thus, we separate the data into M classes in an unsupervised manner, but then for the evaluation, we consider the labeled data to associate each group to an actual class. This is a common way to evaluate unsupervised learning for a classification task [4]; more insights are discussed at the clustering step.

1) FEATURE EXTRACTION

In the scenario of the unavailability of any annotated data, the derived pre-trained model is used to extract features. In this case, we consider the learned neural network as a function $f : \mathbb{R}^A \rightarrow \mathbb{R}^E$, which maps each image $x_i \in X$ from its original space $\mathbb{R}^A$ to the representation space $\mathbb{R}^E$. The choice of representation and its related feature vector length E is studied in more detail in subsection V-A1.

2) CLUSTERING

Each obtained representation, $f(x_i)$, from the previous step is used as an input to a clustering algorithm. Since the main focus of this work is on the representation learning part, two off-the-shelf standard clustering algorithms, k-means [35] and spherical k-means [36], are utilized. In k-means, the centroids $\{\mu_1, \ldots, \mu_M\}$ of the cluster sets $C = \{c_m, m = 1, \ldots, M\}$ are found through minimizing the Euclidean distance between each obtained document image representation, $f(x_i)$, and the nearest centroid, μ_m, over all the M clusters using the following objective function $J(C)$:

$$J(C) = \sum_{m=1}^{M} \sum_{f(x_i) \in c_m} \| f(x_i) - \mu_m \|^2. \quad (2)$$

The spherical k-means algorithm is also based on a similar loss, but the cosine similarity is used instead of the Euclidean distance.

Once the cluster centroids $\{\mu_1, \ldots, \mu_M\}$ are obtained during the training process, each test sample is then assigned to its nearest centroid in the unsupervised classification process. Afterwards, each cluster of test samples is assigned to an actual class (i.e., from the test set true labels) in an optimal way using the Hungarian algorithm [37]. This algorithm considers a matching matrix of the predicted cluster labels and true labels and returns the indices of the best matching pairs.

D. SUPERVISED CLASSIFICATION STAGE

If a limited amount of annotated data is available, the learned parameters θ of the same network architecture of the unsupervised pre-training are used as an initialization to improve the supervised classification performance. As illustrated in Fig. 4, this neural network is then fine-tuned on the provided small annotated data with cross-entropy loss function and an M-way softmax classification layer. Notice that M is now the real number of classes of the task.

IV. EXPERIMENTAL SETUP

In this section, the used datasets and the implementation details for the different experiments are explained.

A. DATASETS

During the unsupervised pre-training stage, two datasets have been utilized with our proposed framework. In both datasets,

only the document images are used. The first dataset is Tobacco-3482[2] [4], which contains 3,482 document images and 10 document classes. The second dataset is RVL-CDIP dataset[3] [6]. This dataset originally contains 400,000 document images and 16 document classes, but only a small subset of those images (i.e., up to 5000) has been used throughout the pre-training stage. This is because more images did not further improve the results and made the training longer, as demonstrated at section V-B1.

At the later stage of both unsupervised and supervised classification, only the Tobacco-3482 dataset has been utilized. At the unsupervised classification stage, the document images are used solely without their associated labels; while at the supervised classification stage, the document images of the Tobacco-3482 dataset and their related labels have been utilized.

At the unsupervised pre-training stage and when using the Tobacco-3482 dataset, we performed the process ten times, one for each partition using 1,000 samples of the related training set. This is to guarantee that all the test samples, at the later classification stages, are completely unseen and have not been used previously during pre-training. On the other hand, when using RVL-CDIP dataset for pre-training, we performed the process only once since all the used samples are considered unseen for the testing process.

To evaluate our document image classification approach at either the unsupervised classification or the supervised classification stages, we follow the same evaluation protocol presented in the literature [5]–[7] to guarantee fair comparisons. Initially, the Tobacco-3482 dataset is divided into 1,000 samples for training and the rest of the samples (2,482) for testing. Since the samples in the original dataset are unevenly distributed between its 10 classes, we make sure that the training set contains exactly 100 samples per class. Then, the training set is divided into 800 samples for training and 200 for validation, where each class is represented with 80 images for training and 20 images for validation. To guarantee a reliable estimation of the proposed approach performance, we report the median classification accuracy of ten randomly-created partitions of the dataset.

B. IMPLEMENTATION DETAILS

All the provided results are based on implementations carried out on an Nvidia GeForce GTX 960 GPU using Theano [38] and Keras API.[4]

For the pre-training stage, the Adam optimization algorithm [39] has been used to train our models with a learning rate of $1e-4$ for 120 epochs; while at the supervised classification stage, the same algorithm has been utilized with $1e-6$ learning rate for 1100 epochs.

During the pre-training stage, the unlabeled training data has been subdivided into batches of 5 samples, where for each epoch, the run-time was around 2 seconds per batch. While during the supervised classification stage, the run-time was around 8 seconds per epoch using 800 samples for training and 200 samples for validation.

At the unsupervised classification stage, the number of times the clustering algorithm will be run with randomly initialized centroids (*'n_init'*) and the maximum number of iterations for each run (*'max_iter'*) are set to 50 and 300, respectively, in case of k-means; while they are set to 150 and 300 in case of spherical k-means. In addition, the 'linear_assignment' function provided by scikit-learn library [40] is used to implement the Hungarian algorithm.

V. RESULTS AND DISCUSSION

A. UNSUPERVISED FEATURE LEARNING

In this subsection, we discuss in details the unsupervised classification performance and the effect of the learned representation on it.

1) SELECTION OF THE LEARNED REPRESENTATION

To study the effect of the learned representation on the unsupervised classification performance, various experiments have been performed using a partition of the Tobacco-3482 dataset. To evaluate the unsupervised classification performance (i.e. clustering is well-matched with the test set's true labels), we follow the literature [4] in computing the purity [41] and the Adjusted Rand Index (ARI) [42].

Specifically, at the feature extraction stage, we study the correlation between the different characteristics of the learned document representations and the unsupervised classification (clustering) performance. The representation characteristics mentioned here refer to the location of the layer to extract the features from and its associated feature vector length E, table 1 provides more details about the different types of layers and their associated locations in the neural network and related output shapes.

Table 2 shows the performance of different learned representations with various locations and dimensionality that ranges from $E = 4,096$ to $E = 43,264$. Although the *flatten* representation has a larger feature vector ($E = 9,216$) than the *dense_2* representation ($E = 4,096$), the former performs better than the latter. This is due to the fact that the *flatten* representation preserves the spatial locality information of its obtained features unlike *dense_2*. On the other hand, since the number of the unlabeled training samples is limited ($N = 1000$), and considering the curse of dimensionality, it is understandable that both high-dimensional representations *conv_5* ($E = 43,264$) and *flatten+dense_1* ($E = 13,312$) obtain a poor performance despite preserving full/some spatial locality information about their features.

[2] https://lampsrv02.umiacs.umd.edu/projdb/project.php?id=72
[3] http://scs.ryerson.ca/ aharley/rvl-cdip/
[4] https://github.com/keras-team/keras

TABLE 2. The unsupervised classification (clustering) ARI and purity when utilizing various learned representations (i.e., on a partition of the Tobacco-3482 dataset).

Representation	Feature vector length (E)	k-means		Spherical k-means	
		ARI	Purity	ARI	Purity
conv_5	43,264	0.1387	0.4057	0.2321	0.4899
flatten + dense_1	13,312	0.2094	0.4694	0.2742	0.5254
flatten	9,216	0.2726	0.5242	**0.2759**	**0.5294**
dense_2	4,096	0.2141	0.4895	0.2194	0.5153

TABLE 3. The unsupervised classification (clustering) ARI and purity results of our learned representation and the state-of-the-art representations.

Representation	ARI	Purity
G-BOW-RF [4]	0.21	0.48
SP-RF [4]	0.22	0.46
HVP-E [4]	0.18	0.46
HVP-RF [4]	0.24	0.49
Proposed U-FL (w/o add. data) -k-means-	0.27	0.52
Proposed U-FL (w/ add. data) -k-means-	**0.29**	**0.54**
Proposed U-FL (w/o add. data) -spherical k-means-	0.28	0.53
Proposed U-FL (w/ add. data) -spherical k-means-	0.27	0.52

TABLE 4. The supervised classification median and mean accuracy, on the Tobacco-3482 dataset, with different parameters initialization methods.

Parameters initialization method	Median accuracy (%)	Mean accuracy±std (%)
No U-PT	63.38	62.74±0.017
Proposed U-PT (w/o add. data)	65.01	65.13±0.012
Proposed U-PT (w/ add. data)	68.86	68.95±0.012

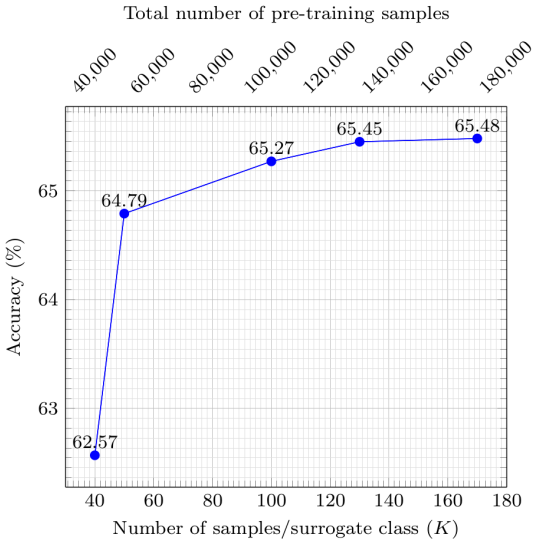


FIGURE 5. The supervised classification accuracy, on a partition of the Tobacco-3482 dataset, with different numbers of used samples/surrogate class (K) and fixed 1000 surrogate classes (N).

Our best results are obtained when using the *flatten* representation for both clustering algorithms, k-means and spherical k-means.

2) UNSUPERVISED CLASSIFICATION (CLUSTERING) RESULTS

Table 3 reports the unsupervised classification results using our proposed unsupervised feature learning (U-FL) based representations, which show an improvement in the performance compared to the best four performing representations in the literature [4]. These codewords based representations are either global-based (G-BOW) or partitioning-based that use either spatial-pyramid (SP) or horizontal vertical partitioning (HVP) to capture the spatial dependencies. Afterward either Euclidean distance (E) or random forest (RF) is used to compute similarities. For our proposed approach, we compare two configurations: 'without additional data (w/o add. data)' refers to using 1000 training samples (unlabeled) of Tobacco-3482 at pre-training, while 'with additional data (w/ add. data)' denotes utilizing 3000 unlabeled samples from RVL-CDIP dataset. In our experiments, the best configuration seems to be k-means with additional data although the difference with respect to the other configurations of our algorithm is relatively small.

Compare with previous approaches, our proposed representation outperforms the HVP-RF representation [4] by 4 points, in both ARI and purity, without the need of any additional data (U-FL (w/o add. data) -spherical k-means-) and 5 points using additional data (i.e., 3000 unlabeled samples from RVL-CDIP dataset) (U-FL (w/ add. data) -k-means-).

B. UNSUPERVISED PRE-TRAINING

This subsection studies the supervised classification performance and its correlation with the pre-training parameters.

1) SELECTION OF THE PRE-TRAINING PARAMETERS

We study the importance of the number of surrogate classes N and the number of samples per surrogate class K on the supervised classification task using a partition of the Tobacco-3482 dataset for evaluation.

First, we study the correlation between the supervised classification performance and the used number of samples per surrogate class K. To do so, we examine the classification performance with various K values using the Tobacco-3482 dataset at the unsupervised pre-training stage with 1000 surrogate classes ($N = 1000$). Fig. 5 shows that increasing the number of samples per surrogate class K results in an improvement in the accuracy that saturates as the number of samples becomes larger.

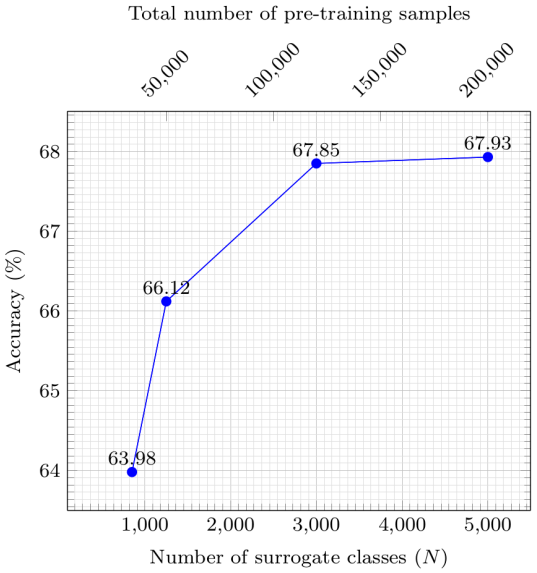


FIGURE 6. The supervised classification accuracy, on a partition of the Tobacco-3482 dataset, with different numbers of utilized surrogate classes (N) and a fixed 40 samples per surrogate class (K).

Then, to examine the correlation between the supervised classification performance and the number of used surrogate classes N (i.e., and consequently the total number of pre-training samples), we apply our proposed approach with various N values. This is performed using the RVL-CDIP dataset with 40 samples per surrogate class ($K = 40$). Note that, in this experiment, the RVL-CDIP dataset is used instead of the Tobacco-3482 (only at the unsupervised pre-training stage), where it offers studying the performance with surrogate classes N values that are beyond 1000 (i.e., the Tobacco-3482 dataset is limited to 1000 training samples). On the other hand, the Tobacco-3482 is still used at the supervised classification stage.

Fig. 6 shows that the accuracy generally improves when increasing the number of used surrogate classes N with a clear saturation after a certain point. For instance, in the last point of Fig. 6, although the number of surrogate classes N has been increased by 2000 (i.e., 66%), the classification performance has not improved significantly, only by 0.08%. This is expected since utilizing more surrogate classes can lead to considering too similar images as different classes, which leads to harder pre-training discrimination and less effective learned parameters [32].

2) SUPERVISED CLASSIFICATION RESULTS

Table 4 demonstrates the supervised classification median and mean accuracy on the Tobacco-3482 dataset, where the parameters initialization is with either i) no pre-training, ii) our proposed unsupervised pre-training (U-PT) based learned parameters θ without any additional data (w/o add. data) (i.e., based on the training data of the Tobacco-3482 dataset using 1000 surrogate classes N with 100 samples/class K (100K samples)), or iii) our proposed unsupervised pre-training (U-PT) based learned parameters θ with additional related unlabeled data (w/ add. data) (i.e., based on a small portion of the training data of the RVL-CDIP dataset using 3000 surrogate classes N with 40 samples/class K (120K samples)). Note that the used N and K values are based on a trade-off between the accuracy and the computational cost of the algorithm (i.e., values where the accuracy starts to saturate while the computation is still moderate).

The obtained results show that incorporating our proposed unsupervised pre-training (U-PT) based learned parameters θ can efficiently and consistently lead to a boost in the supervised classification accuracy over the performance of the method when trained from scratch. The improvement is over 1.5% without the need of any extra data and using only an unlabeled version of the same training data to be used at the supervised classification stage. Additionally, our approach is capable of boosting the classification accuracy to over 5% when substituting the previously used data with unlabeled data from a related dataset (e.g., RVL-CDIP dataset).

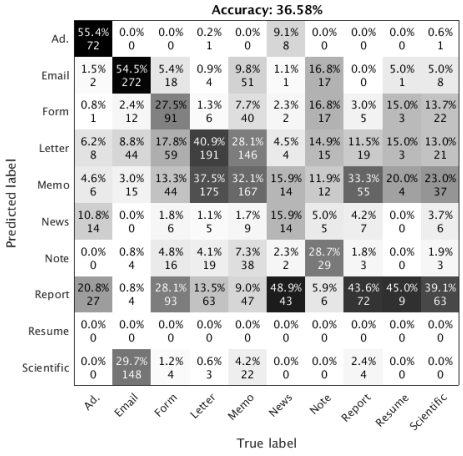
C. DISCUSSION

To illustrate the performance of both unsupervised and supervised classification on the same metric space, the accuracy of the unsupervised classification process is calculated through efficiently utilizing the Hungarian algorithm [37] to find the optimal assignment between each cluster of document images and its corresponding class in the ground truth (true label).

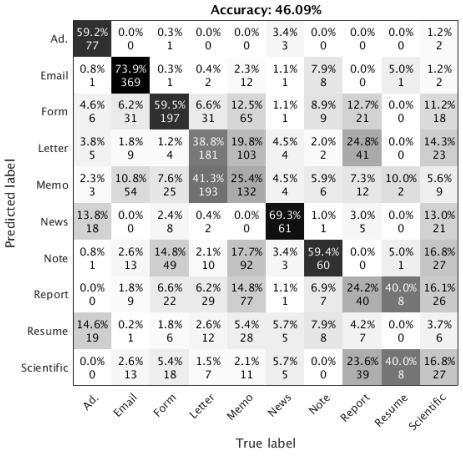
Fig. 7 and table 5 demonstrate the impact of utilizing the learned pre-trained model on the document image classification performance with both of its unsupervised and supervised settings, specifically: i) on the unsupervised classification accuracy using the model's learned representations ii) on the supervised classification accuracy using the model's learned parameters θ. In both cases, the results are compared to their relevant baselines.

Fig. 7 reports the confusion matrices of tests performed on one partition of the Tobacco-3482 dataset. We observe that incorporating our proposed unsupervised feature learning (U-FL) based representations with the unsupervised classification leads to a better class grouping. This is except for some classes which have low inter-class layout variations with each other (i.e., the high layout similarities between the classes of report, resume and scientific). Similarly, for the supervised classification, our proposed unsupervised pre-training (U-PT) based learned parameters θ yields better grouping results in many classes comparing to training the network from scratch.

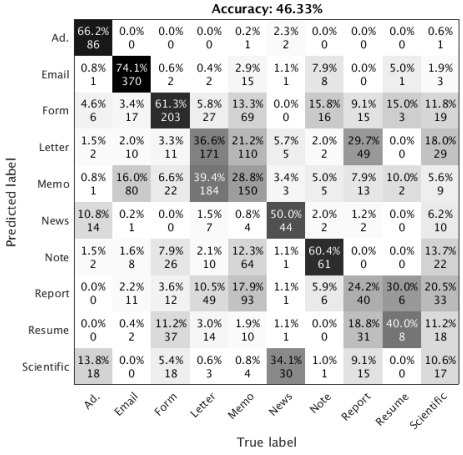
Table 5 compares the performance of our methods for supervised and unsupervised classification and other approaches. In order to have a fair comparison, all methods are trained (either supervised or unsupervised) on 1000 samples of the Tobacco-3482 dataset. We can separate the methods in unsupervised (upper part of the table) and supervised (lower part of the table). All the supervised methods outperform the unsupervised ones. This is expected as in the unsupervised case, classes are grouped based only on



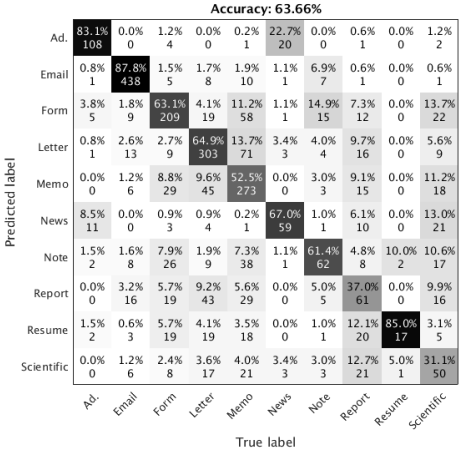
(a) Unsup. classification: No U-FL.



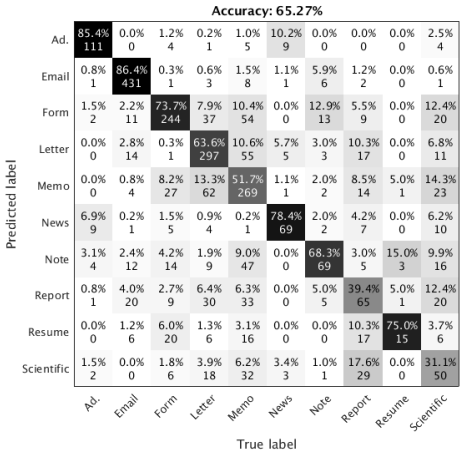
(b) Unsup. classification: Proposed U-FL (w/o add. data).



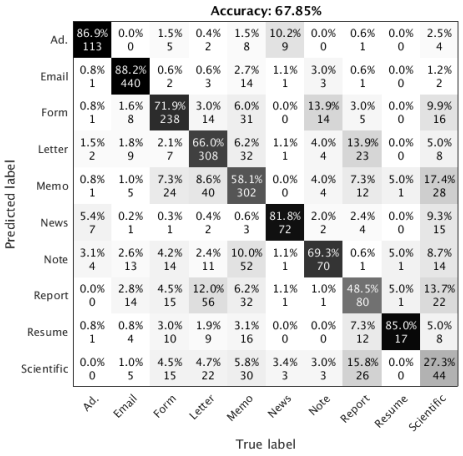
(c) Unsup. classification: Proposed U-FL (w/ add. data).



(d) Supervised classification: No U-PT.



(e) Supervised. classification: Proposed U-PT (w/o add. data).



(f) Supervised. classification: Proposed U-PT (w/ add. data).

FIGURE 7. Confusion matrices for different models on one partition of the Tobacco-3482 dataset. (a) Unsupervised classification using features from a randomly initialized network. (b) Unsupervised classification using features from a network pre-trained on 1000 non-annotated samples. (c) Unsupervised classification using features from a network pre-trained on 3000 non-annotated samples. (d) Supervised classification without any pre-training. (e) Supervised classification with unsupervised pre-training on 1000 non-annotated samples. (f) Supervised classification with unsupervised pre-training on 3000 non-annotated samples.

TABLE 5. Classification median accuracy (i.e., on the Tobacco-3482 dataset-ten partitions-) for unsupervised and supervised methods with different pre-training approaches.

	Method	Pre-training		Median accuracy (%)
		Unsupervised	Supervised	
Unsup.	No U-FL	-	-	36.76
	HVP-RF [21]	-	-	43.80
	Proposed U-FL (w/o add. data)	1000	-	45.26
	Proposed U-FL (w/ add. data)	3000	-	46.25
Supervised	No U-PT	-	-	63.38
	Proposed U-PT (w/o add. data)	1000	-	65.01
	Proposed U-PT (w/ add. data)	3000	-	68.86
	S-PT (w/ ImageNet)	-	$\sim 1,000,0000$	72.89
	S-PT (w/ document images) [5]	-	$320,0000$	90.04

clustering approaches and no labels are used. Among the unsupervised methods, we can see that the features extracted from our network architecture without any pre-training (No U-FL) perform quite poorly. However, when we use the features from our pre-trained network (U-FL), the results are much better. This shows that our unsupervised pre-training approach is very effective in learning good features. Additionally, our methods obtain better results than [21], which is based on a random forest and hand-crafted features that are selected for the specific task. For the supervised classification (lower part of table 5), we can see a similar pattern in which using unsupervised pre-training (U-PT) helps to improve the performance, going from 63.38% to 68.86% for the pre-training with 3000 images. In fact, our unsupervised pre-training (U-PT) gets closer to the performance of a model pre-trained with over one million labelled data (ImageNet). Finally, we see that in the case of having access to a large amount of similar labelled data (e.g., utilizing 320,000 annotated document images in [5]), results can be further boosted up to 90%. Overall, we can see that with limited training data (1000 training samples) and without a proper pre-training (No U-FL and No U-PT), CNN-based methods perform quite poorly. However, incorporating our proposed unsupervised pre-training enables these methods to be trained more effectively and leads to better results without the need of extra annotated data.

VI. CONCLUSION

Contrary to conventional document image classification methods that use either hand-crafted features or supervised pre-training approaches, we propose a visual features learning approach that is based on unsupervised pre-training. The proposed approach uses only unlabeled data to learn a pre-trained model, which is used later for unsupervised and supervised classification. Our approach improves the performance of the document image classification problem in the cases of i) the unavailability of any labeled data, ii) the availability of limited labeled data and iii) the availability of additional unlabeled data. Our experimental results corroborate the capability of our approach to improve the accuracy of CNN-based classification methods. Although other supervised pre-training approaches may provide more improvement in the classification performance, our approach has a crucial advantage of not requiring any additional manually annotated data.

ACKNOWLEDGMENT

The authors thank the NSERC of Canada (Grants no. RGPIN 2014-04649 and RGPIN 2018-04825) for their financial support.

REFERENCES

[1] A. Dengel and F. Dubiel, "Clustering and classification of document structure-a machine learning approach," in *Proc. 3rd Int. Conf. Document Anal. Recognit.*, vol. 2, Aug. 1995, pp. 587–591.

[2] S. S. Bukhari and A. Dengel, "Visual appearance based document classification methods: Performance evaluation and benchmarking," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Aug. 2015, pp. 981–985.

[3] S. Chen, Y. He, J. Sun, and S. Naoi, "Structured document classification by matching local salient features," in *Proc. 21st Int. Conf. Pattern Recognit. (ICPR)*, Nov. 2012, pp. 653–656.

[4] J. Kumar and D. Doermann, "Unsupervised classification of structurally similar document images," in *Proc. 12th Int. Conf. Document Anal. Recognit.*, Aug. 2013, pp. 1225–1229.

[5] M. Z. Afzal, A. Kölsch, S. Ahmed, and M. Liwicki, "Cutting the error by half: Investigation of very deep CNN and advanced training strategies for document image classification," 2017, *arXiv:1704.03557*. [Online]. Available: https://arxiv.org/abs/1704.03557

[6] A. W. Harley, A. Ufkes, and K. G. Derpanis, "Evaluation of deep convolutional nets for document image classification and retrieval," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Aug. 2015, pp. 991–995.

[7] L. Kang, J. Kumar, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for document image classification," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 3168–3172.

[8] A. Das, S. Roy, U. Bhattacharya, and S. K. Parui, "Document image classification with intra-domain transfer learning and stacked generalization of deep convolutional neural networks," 2018, *arXiv:1801.09321*. [Online]. Available: https://arxiv.org/abs/1801.09321

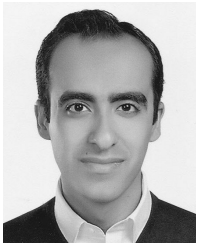
[9] M. Z. Afzal, S. Capobianco, M. I. Malik, S. Marinai, T. M. Breuel, A. Dengel, and M. Liwicki, "Deepdocclassifier: Document classification with deep convolutional neural network," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Aug. 2015, pp. 1111–1115.

[10] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning*, vol. 1. Cambridge, MA, USA: MIT Press, 2016.

[11] S. Abuelwafa, M. Mhiri, R. Hedjam, S. Zhalehpour, A. Piper, C. Wellmon, and M. Cheriet, "Feature learning for footnote-based document image classification," in *Proc. Int. Conf. Image Anal. Recognit.* Cham, Switzerland: Springer, 2017, pp. 643–650.

[12] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?" *J. Mach. Learn. Res.*, vol. 11, pp. 625–660, Feb. 2010.

[13] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.
[14] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.
[15] N. Chen and D. Blostein, "A survey of document image classification: Problem statement, classifier architecture and performance evaluation," *Int. J. Document Anal. Recognit.*, vol. 10, no. 1, pp. 1–16, Jun. 2007.
[16] B. Tang, H. He, P. M. Baggenstoss, and S. Kay, "A Bayesian classification approach using class-specific features for text categorization," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 6, pp. 1602–1606, Jun. 2016.
[17] L. Noce, I. Gallo, A. Zamberletti, and A. Calefati, "Embedded textual content for document image classification with convolutional neural networks," in *Proc. ACM Symp. Document Eng.*, Sep. 2016, pp. 165–173.
[18] J. Kumar, "Efficient machine learning methods for document image analysis," Ph.D. dissertation, Fac. Graduate School, Univ. Maryland, College Park, College Park, MD, USA, 2013.
[19] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[20] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 404–417.
[21] J. Kumar, P. Ye, and D. Doermann, "Structural similarity for document image classification and retrieval," *Pattern Recognit. Lett.*, vol. 43, pp. 119–126, Jul. 2014.
[22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[23] S. Roy, A. Das, and U. Bhattacharya, "Generalized stacking of layerwise-trained deep convolutional neural networks for document image classification," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 1273–1278.
[24] C. Tensmeyer and T. Martinez, "CONFIRM—Clustering of noisy form images using robust matching," *Pattern Recognit.*, vol. 87, pp. 1–16, Mar. 2019.
[25] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.
[26] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, no. 12, pp. 3371–3408, Dec. 2010.
[27] A. Makhzani and B. Frey, "K-sparse autoencoders," 2013, *arXiv:1312.5663*. [Online]. Available: https://arxiv.org/abs/1312.5663
[28] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*. [Online]. Available: https://arxiv.org/abs/1312.6114
[29] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1422–1430.
[30] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 69–84.
[31] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," 2018, *arXiv:1803.07728*. [Online]. Available: https://arxiv.org/abs/1803.07728
[32] A. Dosovitskiy, P. Fischer, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative unsupervised feature learning with exemplar convolutional neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1734–1747, Sep. 2016.
[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[34] J. Wang and L. Perez, "The effectiveness of data augmentation in image classification using deep learning," *Convolutional Neural Netw. Vis. Recognit*, 2017.
[35] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, 2010.
[36] I. S. Dhillon and D. S. Modha, "Concept decompositions for large sparse text data using clustering," *Mach. Learn.*, vol. 42, no. 1, pp. 143–175, 2001.
[37] H. W. Kuhn, "The Hungarian method for the assignment problem," *Nav. Res. Logistics Quart.*, vol. 2, nos. 1–2, pp. 83–97, Mar. 1955.
[38] T. D. Team *et al.*, "Theano: A Python framework for fast computation of mathematical expressions," May 2016, *arXiv:1605.02688*. [Online]. Available: https://arxiv.org/abs/1605.02688
[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980
[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," 2011, *arXiv:1201.0490*. [Online]. Available: https://arxiv.org/abs/1201.0490
[41] C. Manning, P. Raghavan, and H. Schütze, "Introduction to information retrieval," *Natural Lang. Eng.*, vol. 16, no. 1, pp. 100–103, 2010.
[42] L. Hubert and P. Arabie, "Comparing partitions," *J. Classification*, vol. 2, no. 1, pp. 193–218, 1985.



SHERIF ABUELWAFA received the B.Sc. degree in electronics and telecommunications engineering from Alexandria University, Egypt, in 2009, and the master's degree in embedded systems design from Università della Svizzera Italiana (USI), Switzerland, in 2012. He is currently pursuing the Ph.D. degree with the École de technologie supérieure, University of Quebec, Montreal, Canada. From 2012 to 2014, he was a Research and Development Engineer at startups in Switzerland and Germany. His research interests include representation learning, unsupervised feature learning, and document image understanding.

MARCO PEDERSOLI received the B.Sc. degree from the Electronic Engineering Department, University of Brescia, Italy, in 2005, and the M.Sc. and Ph.D. degrees from the Computer Science Department, Autonomous University of Barcelona, Spain, in 2008 and 2012, respectively. From 2012 to 2015, he received the Postdoctoral Fellowship from KU Leuven, Belgium. From 2015 to 2016, he held a postdoctoral position with INRIA Grenoble. Since February 2017, he has been an Assistant Professor with the École de technologie supérieure, University of Quebec, Montreal, Canada. His research interests include the localization and classification of visual object categories for image understanding, weakly supervised learning, and unsupervised learning.

MOHAMED CHERIET (SM'95) received the M.Sc. and Ph.D. degrees in computer science from the University of Pierre et Marie Curie (Paris VI), in 1985 and 1988, respectively. Since 1992, he has been a Professor with the Automation Engineering Department, École de technologie supérieure, University of Quebec, Montreal, Canada, and also a Full Professor, in 1998. He has also been the Founder and Director of the Synchromedia Laboratory, since 1998. He is currently an Expert in computational intelligence, pattern recognition, mathematical modeling for image processing, and cognitive and machine learning approaches and perception. He has authored more than 350 technical articles in the field. He is a Fellow of the International Association for Pattern Recognition and the Canadian Academy of Engineering. He was a recipient of the 2016 IEEE J. M. Ham Outstanding Engineering Educator Award and the 2012 Queen Elizabeth II Diamond Jubilee Medal. He is the Founder and the Former Chair of the IEEE Montreal Chapter of Computational Intelligent Systems. He serves on the Editorial Boards of several renowned journals and international conferences.