IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

DocXClassifier: Towards an Interpretable Deep Convolutional Neural Network for Document Image Classification

SAIFULLAH[1][2], STEFAN AGNE[1][3], ANDREAS DENGEL[1][2], AND SHERAZ AHMED[1][3]
[1]German Research Center for Artificial Intelligence (DFKI) 67663 Kaiserslautern, Germany
e-mail: {saifullah.saifullah,stefan.agne,andreas.dengel,sheraz.ahmed}@dfki.de
[2]TU Kaiserslautern, 67663 Kaiserslautern, Germany
[3]DeepReader GmbH, 67663 Kaiserlautern, Germany

Corresponding author: Saifullah (e-mail: saifullah.saifullah@dfki.de).

ABSTRACT Convolutional Neural Networks (ConvNets) have been thoroughly researched for document image classification and are known for their exceptional performance in unimodal image-based document classification. Recently, however, there has been a sudden shift in the field towards multimodal approaches that simultaneously learn from the visual and textual features of the documents. While this has led to significant advances in the field, it has also led to a waning interest in improving pure ConvNets-based approaches. This is not desirable, as many of the multimodal approaches still use ConvNets as their visual backbone, and thus improving ConvNets is essential to improving these approaches. In this paper, we present DocXClassifier, a ConvNet-based approach that, using state-of-the-art model design patterns together with modern data augmentation and training strategies, not only achieves significant performance improvements in image-based document classification, but also outperforms some of the recently proposed multimodal approaches. Moreover, DocXClassifier is capable of generating transformer-like attention maps, which makes it inherently interpretable, a property not found in previous image-based classification models. Our approach achieves a new peak performance in image-based classification on two popular document datasets, namely RVL-CDIP and Tobacco3482, with a top-1 classification accuracy of 94.17% and 95.57% on the two datasets, respectively. Moreover, it sets a new record for the highest image-based classification accuracy of 90.14% on Tobacco3482 without transfer learning from RVL-CDIP. Finally, our proposed model may serve as a powerful visual backbone for future multimodal approaches, by providing much richer visual features than existing counterparts.

INDEX TERMS Document Image Classification, Modern Convolutional Neural Networks, Modern Training Strategies, Explainable Document Classification, Model Interpretability

I. INTRODUCTION

In this era of digitization, many organizations seek to implement paperless business workflows in their environments, and therefore great emphasis is being placed on intelligent document processing pipelines that are not only capable of automatically digitizing and managing document data, but also extracting various types of information from them. An important step that is fundamental to such document processing pipelines is the early classification of document images, which not only enables efficient document search and retrieval [1], [2], but also helps to improve the performance of downstream processing tasks such as optical character recognition (OCR), key information extraction, and layout analysis [3]. However, the problem of classifying document images is not trivial and proves to be particularly difficult due to the relatively high intraclass variance and interclass similarity. That is, two documents of the same class may look very different, while two documents from different classes may look similar. Nevertheless, many techniques have been proposed in the past to solve this problem. Previous attempts ranged from traditional computer vision-based techniques [4]–[6] to classical machine learning approaches [7], [8]. However, most of these techniques were only applicable to structured document data with relatively low intraclass variance. It was not until the advent of ConvNets [9], [10] that significant breakthroughs were made in this field. ConvNets, with their exceptional feature learning capabilities, not only significantly outperformed conventional techniques on structured data [9], but also achieved superhuman performance on highly unstructured document data [11].

Although ConvNets alone have shown great potential in the field of document image classification, the recent success of Transformers in natural language processing (NLP) has led to a sudden paradigm shift in the field, with more and more emphasis being laid on multimodal techniques. Multimodal techniques attempt to accomplish the task of document classification by integrating the textual, visual, and layout features of the documents and have shown significant performance improvements in recent years [12]–[14]. However these techniques have their own drawbacks. For example, these techniques always require a preprocessing step that uses a standalone OCR software to extract the textual information from the documents [12], [14], and are therefore heavily dependent on the performance and computational overhead of the OCR software. In addition, these approaches typically feed the textual and visual data into either separate, independent streams of deep networks or large Transformers, increasing both the complexity and size of the models. In particular, Transformer based multimodal techniques [14]–[16] also often require extensive pre-training before they can achieve sufficient performance improvements, which in itself can be a costly process. ConvNets, on the other hand, are simple in design, operate independently, and are often much easier to train compared to Transformers. Moreover, ConvNets are an essential component of most state-of-the-art two-stream multimodal approaches [12], [13], [17], where they are used as a backbone to generate visual features and thus, improving the performance of ConvNets is critical to improving the performance of two-stream multimodal approaches.

There have been numerous recent advances in the field of deep learning, such as Transformer-inspired model designs [18], [19], complex data augmentation techniques [20]–[22], and advanced training strategies [23] that have led to improvements in the performance, robustness, and overall generalization capabilities of ConvNets. However, many of these techniques have not been adequately explored in the context of document image classification. In this work, we therefore investigate them with the goal of improving both the performance and interpretability of ConvNets in document image classification, so that they can be used not only as an independent classifier, but also as a better visual backbone for improving future multimodal approaches.

The contributions of this paper can be summarized as follows. We explore the potential of recently proposed ConvNeXt models [18], Learned Aggregation Layer [19], data augmentation strategies such as CutMix [21] and Mixup [20], and training strategies such as Label Smoothing [24] and Exponential Moving Average (EMA) for document image classification, and introduce a ConvNet-based model that not only achieves a new level of excellence in image-based document classification, but also outperforms some of the existing multimodal approaches. Moreover, our proposed models are inherently interpretable by their ability to generate Transformer-like attention maps that can be used to interpret the model's predictions. We evaluate our approach on the two well-known document benchmark datasets, RVL-CDIP and Tobacco3482. On RVL-CDIP, our approach achieves an accuracy of 94.17%, significantly outperforming the previous state-of-the-art, which had an accuracy of 92.31%. On the Tobacco3482 dataset, we train our models with and without RVL-CDIP pre-training and achieve accuracies of 95.57% and 90.14%, outperforming previous state-of-the-art methods that achieved accuracies of 94.04% and 85.9%.

The paper is organized as follows. In the following section, we present a comprehensive overview of the existing literature on document image classification, as well as several new techniques that have recently emerged in the field of Deep Learning. In Section. III, we present the details of our proposed model architectures and the various data preprocessing, data augmentation, and training techniques that we have explored in our work. In Section. IV, we present the results in which we evaluate our proposed techniques on the two document datasets, perform an ablation study on our approach, perform a runtime evaluation, and finally visualize the attention maps that can be generated using our approach to interpret the results of our models. Finally, we present the conclusion and future work in Section. V.

II. RELATED WORK

A. DOCUMENT IMAGE CLASSIFICATION

The topic of document classification has been extensively explored in the past. Earlier attempts to classify document images were mainly based on traditional computer vision techniques, such as exploiting the structural similarity constraints [25] or distinguishing between different document classes based on feature matching [5], [26]. Several classical machine learning approaches, such as K-Nearest Neighbors [7], Random Forest Classifiers [26] and Hidden Markov Models [8] have also been proposed in the past. For a detailed overview of the classical approaches, we refer the reader to a related survey [27].

With the advent of deep learning, the field of document image classification experienced a major performance boost. Kang *et al.* (2014) [9] were the first to investigate ConvNets in the context of document image classification and were able to achieve significant performance improvements over classical feature engineering approaches even with a very simple, shallow network. Harley *et al.* (2015) [11] presented the first large-scale benchmark dataset RVL-CDIP for document image classification and investigated the use of deeper ConvNets in their work, showing significant performance improvements. Similarly, Afzal *et al.* (2015) [10] explored the potential of transfer learning in combination with deep networks in their work, showing that fine-tuning models already pre-trained on the large-scale ImageNet [28] dataset can lead to significantly better feature representations and consequently better performance. Afzal *et al.* (2017) [29]

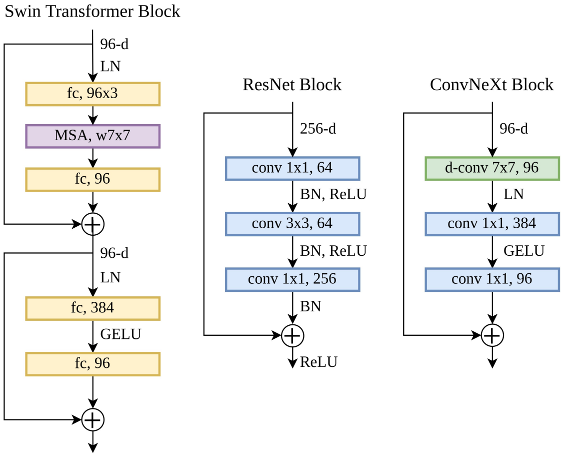


FIGURE 1. Block configurations of ConvNext, ResNet and Swin Transformer are shown for comparison.

later extended their work to much deeper ConvNets achieving breakthrough performance improvements in document image classification. In a more recent approach, Ferrando *et al.* (2020) [30] investigated parallel training techniques on EfficientNet [31] models and achieved a new peak performance for image-based document classification. Due to their recent success in classification of natural images, Vision Transformers (ViTs) [32] have also gained some attention in document image classification [33], however, more work is needed before they can match the performance of the latest ConvNets.

Recently, there has been an increased emphasis on multimodal classification techniques [13], [34], [35], in which document images are preprocessed to extract the textual content using stand-alone OCR software, and then visual, textual, and other layout features are used together for classification. Initial work in this area focused mainly on generating textual and visual embeddings using two separate deep network streams [12], [13] and then integrating them into a single embedding for final classification. Transformer-based multimodal techniques have also become more popular recently. Xu *et al.* recently proposed the LayoutLM [14] and LayoutLMV2 [15] models, which are large-scale Transformer networks that simultaneously use visual, textual, and layout features as input and produce an integrated multimodal document representation for document classification. Powalski et al. (2021) [16] went a step further and used two transformer networks, an encoder and a decoder, instead of a single transformer network to perform the task of multimodal document classification. However, both approaches of these approaches require pre-training with large amounts of document data. In a slightly different direction, Graph ConvNets [35] have also been recently explored for multimodal classification and show promising results.

B. MODERN CONVNET DESIGNS AND DATA AUGMENTATIONS

Since the introduction of the groundbreaking AlexNet [36] architecture by Krizhevsky *et al.* in 2012, the field of deep learning has evolved rapidly. Over the years, many different types of ConvNets such as VGG [37], ResNet [38], EfficientNet [31], etc. have been proposed, each focusing on a different aspect such as performance, scalability, and efficiency, which has led to many useful design principles for the research community. However, due to the recent success of ViTs [32], which significantly outperform standard ConvNets in image classification, there is growing interest in more generic Transformer based vision backbones that can be used for a wide range of computer vision tasks. As a result, techniques such as Swin Transformers [39] have recently been proposed that attempt to introduce ConvNets-like generalization capabilities into Transformers. Multi-layer Perceptron (MLP) based models have also been recently proposed to solve the image classification problem [40]–[42]. However, since ConvNets are already well researched for a variety of image processing tasks, others [18], [19] are instead attempting to modernize ConvNets by introducing design changes inspired by Transformers to achieve the performance comparable to Transformers.

In addition to advances in neural network design, many advanced data augmentation strategies have also been proposed in recent years that allow the models to learn better features, reduce model overfitting, and increase overall model performance and robustness. Techniques such as AutoAugment [43], and RandAugment [22] improve model generalization by automatically searching for the optimal data augmentation policy. Random Erasing [44] is a simple augmentation techniques that randomly erases sections of the image. Strategies like CutMix [21] and Mixup [20], on the other hand, attempt to regularize the model by mixing the samples from multiple classes to generate new samples. These techniques are also often accompanied with Label Smoothing [24], which regularizes the models by preventing them from predicting the output labels too confidently.

III. METHODOLOGY

In this section, we describe in detail our proposed ConvNet architecture, the data augmentation techniques, and the training strategies that we have used in our study.

A. MODEL ARCHITECTURE

ConvNeXt was recently proposed by Liu *et al.* [18] as a modernized version of a traditional ConvNet that is not only heavily inspired by the state-of-the-art ViTs, but can also outperform them in image classification. In particular, ConvNeXt was developed by making various design modifications to the standard ResNet model [38]—modifications inspired by both modern ConvNets and the recently introduced Swin Transformers [39], a variant of ViTs. In the following, we briefly explain these modifications, which mainly fall into two main categories: Macro Design and Micro Design.

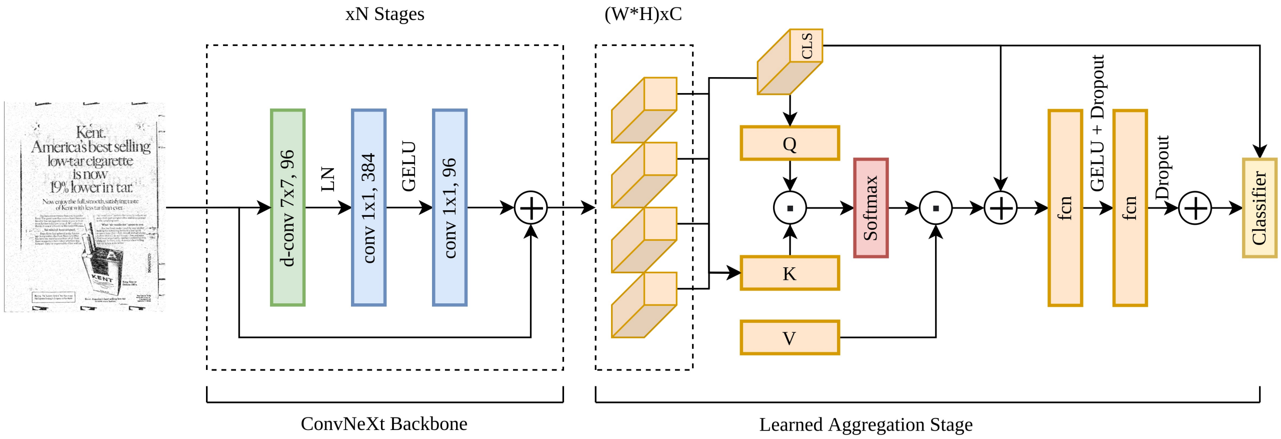


FIGURE 2. Complete configuration of the proposed DocXClassifier model. The base ConvNeXt model is used as the backbone for generating the feature vectors of the image, which are then fed into a Learned Aggregation Layer to generate the attention maps. Finally, a linear classification head is used to generate the class scores.

TABLE 1. Number of channels and blocks per stage for different ConvNeXt variants.

Model	Channels	Blocks
ConvNeXt-T	(96, 192, 384, 768)	(3, 3, 9, 3)
ConvNeXt-S	(96, 192, 384, 768)	(3, 3, 27, 3)
ConvNeXt-B	(128, 256, 512, 1024)	(3, 3, 27, 3)
ConvNeXt-L	(192, 384, 768, 1536)	(3, 3, 27, 3)
ConvNeXt-XL	(256, 512, 1024, 2048)	(3, 3, 27, 3)

Macro Design. The first major design modification was to change the stage compute ratio from 1:1$\frac{1}{3}$:2:1 to 1:1:3:1, directly inspired by the Swin Transformers [39], which have a stage compute ratio of 1:1:9:1. For example, compared to the ResNet-50 model, the blocks per stage in ConvNext were changed from $(3, 4, 6, 3)$ to $(3, 3, 9, 3)$. Another important design change was the replacement of the initial stem cell of the model with a Patchify layer [18], as is common in ViTs [32]. The stem cell in the standard ResNet models contains a 7x7 convolutional layer followed by a max-pooling layer the purpose of which is to downsample the input image to a smaller size. The ConvNeXt models replace this with a Patchify layer [18], implemented with a non-overlapping convolutional layer of kernel size 4x4 and a stride 4. Next, taking inspiration from the ResNext-style grouped convolutions, depth-wise convolutions were introduced into the model design, a special case of grouped convolutions where the number of groups is set equal to the number of channels. In addition, the inverted bottleneck was introduced in each block, but with the convolutional layers shifted up in order, a design decision again inspired by Transformers, where the multi-self-attention blocks are generally placed before the fully connected layers. Finally, the layers were modified to use a 7x7 kernel size instead of a 3x3 kernel size, which proved to be optimal with the newly introduced design decisions.

Micro Design. Some minor architectural changes were also made. For example, the ReLU activations were replaced with GELU activations, which are commonly used in latest Transformers. The total number of activations were reduced so that there was only a single activation function at the end of each block. The total number of normalization layers were also reduced and batch normalization was removed in favor of layer normalization. Finally, the initial residual block in ResNet was removed and instead a separate downsampling layer, followed by layer normalization, was added between each stage to mimic the Swin Transformers.

Although a considerable number of design changes were made, the resulting ConvNeXt model is just another ConvNet without any sophisticated components. A comparison of the design of a single block of ConvNeXt, ResNet, and Swin Transformer is shown in Fig. 1. The different variants of this model are defined by varying the number of channels and the number of blocks in each stage, resulting in the configurations shown in the Table 1. In this study, we investigate the performance of three variants, ConvNeXt-B, ConvNeXt-L, and ConvNeXt-XL, in document image classification. **Attention-Based Pooling.** Since the original ConvNeXt models are simply ConvNets, they are not capable of generating attention maps out-of-the-box. To add this capability, we replaced the global average pooling with attention-based pooling by augmenting the model with a Learned Aggregation Layer on top. The Learned Aggregation Layer is a Transformers-inspired cross-attention layer recently proposed by Tourvan *et al.* [19] that aggregates the output feature vectors generated by ConvNets based on their similarity to a target class vector. In particular, we first reshape the ConvNet feature map with dimensions BxCxHxW to a BxCx(H*W) dimension, resulting in H*W output feature vectors. A query class token (Q) is then used to aggregate the feature map vectors as a weighted summation based on their similarity to a trainable class (CLS) vector. In particular, the similarity

TABLE 2. A comparison of the classification accuracy of different approaches on the RVL-CDIP and Tobacco3482 datasets

Modality	Model	Inference Time [ms]	# of Parameters	Domain-specific pre-training	RVL-CDIP	Tobacco3482 (RVL-CDIP pre-training)	Tobacco3482 (ImageNet pre-training)
Image	Holistic CNN (Harley *et al.*, 2015 [11])	-	-		89.80%	-	-
	AlexNet (Afzal *et al.*, 2017 [29])	1.1	57M		88.60%	90.04%	75.73%
	GoogleNet (Afzal *et al.*, 2017 [29])	1.2	5.6M		89.02%	88.40%	72.98%
	ResNet-50 (Afzal *et al.*, 2017 [29])	1.1	23.5M		90.40%	91.13%	67.93%
	VGG-16 (Afzal *et al.*, 2017 [29])	1.3	134M		90.97%	91.01%	77.52%
	Stacked CNN Single (Das *et al.*, 2018 [45])	-	-		91.11%	-	-
	Stacked CNN Ensemble (Das *et al.*, 2018 [45])	-	-		92.21%	-	-
	EfficientNet (Ferrando *et al.*, 2020 [30])	2.3	17.6M		92.31%	94.04%	85.99%
	DocXClassifier-B/384 (Ours)	6.53	95.4M		**94.00%**	**95.29%**	**87.43%**
	DocXClassifier-L/384 (Ours)	10.0	204M		**94.15%**	**95.57%**	**88.43%**
	DocXClassifier-XL/384 (Ours)	16.1	356M		**94.17%**	**95.43%**	**90.14%**
Multimodal	MobileNetV2+Text (Audebert *et al.*, 2019 [13])	-	-		90.60%	-	87.80%
	EfficientNet + BERT (Ferrando *et al.*, 2020 [30])	-	127.6M		-	94.90%	89.47%
	LadderNet (Sarkhel *et al.*, 2019 [46])	-	-		92.77%	82.78%	-
	Multimodal Ensemble (Dauphinee *et al.*, 2019 [47])	-	-		93.07%	-	-
	Multimodal GCN (Xiong *et al.*, 2021 [35])	-	49M		93.45%	-	-
	LayoutLM$_{\text{BASE}}$ (Xu *et al.*, 2020 [14])	-	160M	✓	94.42%	-	-
	TILT$_{\text{LARGE}}$ (Powalski *et al.*, 2021 [16])	-	780M	✓	95.52%	-	-
	EfficientNet+BERT (Kanchi *et al.*, 2022 [48])	-	197M		95.48%	**95.7%**	**90.3%**
	LayoutLMv2$_{\text{LARGE}}$ (Xu *et al.*, 2021 [15])	-	426M	✓	95.64%	-	-
	NasNet$_{\text{Large}}$+BERT$_{\text{BASE}}$ (Bakkali *et al.*, 2020 [17])	-	197M		**97.05%**	-	-

is computed using the standard attention operation [49], as shown in Fig. 2, where K, Q, and V represent the query, key, and value matrices of an attention block, respectively. The resulting aggregated vector is then added to the CLS vector and processed by a feed-forward network. Finally, a linear classifier is added to the model to perform the classification. The final model configuration with this modification is shown in Fig. 2, which we refer to as DocXClassifier. The complete implementation details of the model can be found at https://github.com/saifullah3396/docxclassifier.git.

B. DATA PREPROCESSING AND AUGMENTATION

In this section, we describe the data preprocessing steps and the augmentation strategies used in our experiments. Basic preprocessing steps include converting grayscale images to RGB color space, downscaling the images to a fixed input resolution of 384x384, and normalizing the images by subtracting the ImageNet mean $(0.485, 0.456, 0.406)$ and dividing by the ImageNet standard deviation $(0.229, 0.224, 0.225)$, as done in previous works [10], [30]. We also explored more advanced data augmentation strategies to improve generalization. In much previous work on document image classification [30], [50], we have encountered the common belief that data augmentation techniques developed for natural images cannot be directly applied to document images due to the fundamental differences between these two image types. As a result, these works have typically performed only minor augmentations to document images, such as a simple shear transformation [30], [50]. In this work, we show that using more aggressive data augmentation techniques typically used for natural images can actually improve the generalization and performance of the networks. The data augmentation techniques we used in our experiments are RandAugment [22], Random Erasing [44], CutMix [21], and Mixup [20]. For RandAugment, we used the implementation from the timm [1] library which randomly applies various data augmentations such as changing image color, brightness, contrast, sharpness and applying transformations such as translation, rotation, shear.

C. IMPLEMENTATION DETAILS

In this section, we provide the details about the training strategies used in each of our experiments.

Training on RVL-CDIP. Since transfer learning has already proven to be successful in the field of document image classification [29], instead of training the models from scratch, we initialized them with the ImageNet-22k [28] pre-trained weights and then fine-tuned them on the RVL-CDIP dataset. All models were trained on 4-8 A100 GPUs with DistributedDataParallel (DDP) using the AdamW optimizer and a cosine decay learning rate strategy with no warm-up period. We chose a base learning rate of 8e-4, corresponding to a batch size of 64, and scaled it linearly with different configurations of batch size, varying between 64, 128, and 256. Since the weights of the learned aggregation stage were initialized from scratch, we found it difficult to train the models DocXClassifier models end-to-end, and therefore we trained them in two steps. First, we fine-tuned the base ConvNeXt models for 30 epochs to achieve the desired classification performance. Then, we froze the weights of the base model, used them to initialize our DocXClassifier variants, and trained only the learned aggregation stage along with the classifier. We also used the regularization techniques Stochastic Depth [23] and Label Smoothing [24] to prevent overfitting of the model, and applied Layer Scale [51] with an initial value of 1e-6. We also used the Exponential Moving Average (EMA) [18] to train all our models, which lead to significant performance improvements in our experiments.

Training on Tobacco3482 On the Tobacco3482 dataset, we trained the models with two different configurations: with

---

[1] https://github.com/rwightman/pytorch-image-models

TABLE 3. Evaluation of the ConvNeXt models with different training settings.

Model	Accuracy (RVL-CDIP)	# of Parameters
ConvNeXt-B/224 (Aug$_{\text{Basic}}$)	92.10%	87.6M
ConvNeXt-B/224 (Aug$_{\text{Basic}}$ + Aug$_{\text{cutmixup}}$)	92.63%	87.6M
ConvNeXt-B-384 (Aug$_{\text{Basic}}$)	93.13%	87.6M
ConvNeXt-B/384 (Aug$_{\text{Basic}}$ + Aug$_{\text{cutmixup}}$)	93.60%	87.6M
ConvNeXt-B/384 (Aug$_{\text{ImageNet}}$)	93.21%	87.6M
ConvNeXt-B/384 (Aug$_{\text{ImageNet}}$ + Aug$_{\text{cutmixup}}$)	93.74%	87.6M
ConvNeXt-L/384 (Aug$_{\text{ImageNet}}$ + Aug$_{\text{cutmixup}}$)	93.75%	196M
ConvNeXt-XL-384 (Aug$_{\text{ImageNet}}$ + Aug$_{\text{cutmixup}}$)	93.81%	348M
ConvNeXt-B/384 (Aug$_{\text{ImageNet}}$ + Aug$_{\text{cutmixup}}$ + EMA)	94.04%	87.6M
ConvNeXt-L/384 (Aug$_{\text{ImageNet}}$ + Aug$_{\text{cutmixup}}$ + EMA)	94.15%	196M
ConvNeXt-XL/384 (Aug$_{\text{ImageNet}}$ + Aug$_{\text{cutmixup}}$ + EMA)	94.17%	348M
DocXClassifier-B/384 (Aug$_{\text{ImageNet}}$ + Aug$_{\text{cutmixup}}$ + EMA)	**94.00%**	95.4M
DocXClassifier-L/384 (Aug$_{\text{ImageNet}}$ + Aug$_{\text{cutmixup}}$ + EMA)	**94.15%**	204M
DocXClassifier-XL/384 (Aug$_{\text{ImageNet}}$ + Aug$_{\text{cutmixup}}$ + EMA)	**94.17%**	356M

RVL-CDIP pre-training and with ImageNet pre-training. In the first configuration, we simply selected the DocXClassifier models that performed best on the RVL-CDIP dataset and further fine-tuned them on the Tobacco3482 dataset. In this case, we used the same training hyperparameters as above, except that we did not apply EMA in this case as it did not seem to yield any improvements. In the second configuration, we followed the same approach as RVL-CDIP, initializing the models with the pre-trained weights from ImageNet-22k [28] and then fine-tuning them directly on the Tobacco3482 dataset in a two-step process. The hyper-parameters used in this configuration were the same as those used in RVL-CDIP training, except for the learning rate and the number of epochs which were set to 5e-5 and 90, respectively.

IV. EXPERIMENTS AND RESULTS

A. DATASETS

To evaluate the performance of our proposed approach on document image classification task, we selected two popular document datasets: RVL-CDIP and Tobacco3482. RVL-CDIP is a large-scale document dataset that has been widely used as a benchmark for document image classification in a number of previous works [11], [14], [29], [30]. The dataset consists of 400K labeled document images with 16 class labels and has training, testing, and validation splits of 320K, 40K, and 40K in size, respectively. Tobacco3482, on the other hand, is a smaller dataset with only 3482 labeled document images, but is still widely popular for the task of document image classification. There is no predefined partitioning for this dataset. Therefore, we prepared the training set by randomly selecting 80% of the samples per class label, resulting in a training and test set of size 2782 and 700, respectively. Since both datasets are subsets of a much larger dataset, there is some overlap between them. Therefore, for all our experiments, we removed the overlapping images from the training set of RVL-CDIP, reducing the size of the training set to 319,756.

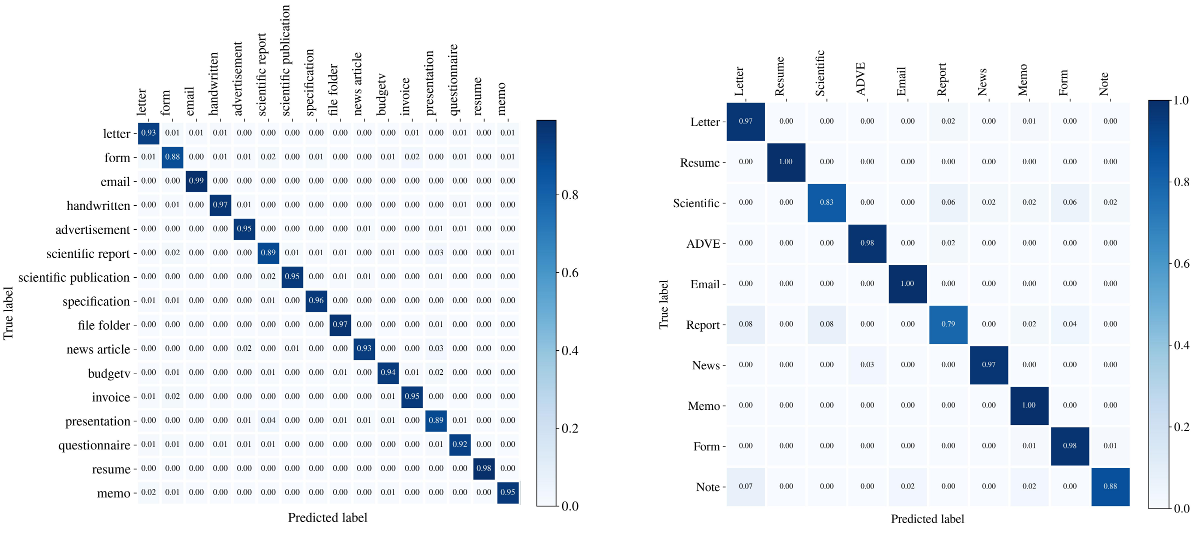
B. OVERALL EVALUATION

Results on RVL-CDIP. Table 2 shows a comparison of the top-1 classification accuracy achieved on the RVL-CDIP and Tobacco-3482 datasets by our approach, previous image-based baseline solutions, and several multimodal approaches that use either text, layout, or both in addition to image data for classification. As can be seen from the table, our best performing model DocXClassifier-XL achieved 94.17% accuracy on the RVL-CDIP dataset, outperforming all previous image-based methods by a significant margin of +1.86%. It is interesting to note that even our lightest variant DocXClassifier-B achieved a comparable accuracy of 94.00%, and performed significantly better than all existing image-based models as well as some of the more sophisticated multimodal approaches [35], [46], [47], thus representing a good trade-off between accuracy and computational cost. It is important to note that two of the best performing multimodal solutions, those of Kanchi *et al.* (2022) [48] and Bakkali *et al.* (2020) [17], simply combined ConvNet-based visual backbones (EfficientNet and NasNet, respectively) with a Transformer-based textual backbone (BERT) to achieve extraordinary improvements in document classification. We suspect that using our improved ConvNet models as visual backbones in such multimodal approaches could lead to even better results.

We also present the confusion matrices of our proposed DocXClassifier-XL model on the two datasets in Fig. 3. As we can see from Fig. 3a, many of the classes are classified correctly to a large extent, but some of the classes are quite strongly confused with the others. For example, the two classes Presentation and Scientific Report have an overlap of 3-4%. This finding is similar to that reported by Kanchi *et al.* (2022) [48, Fig. 9] on their multimodal approach. In contrast to their results, however, our approach performs better in distinguishing between Scientific Report and Scientific Publication classes. Overall, our approach falls short of their multimodal approach especially for the Form, Questionnaire, and Scientific Report classes, from which we can conclude that these three classes must benefit strongly from textual features of the documents.

Results on Tobacco3482. On the Tobacco3482 dataset, we see similar behavior to RVL-CDIP, where the DocXClassifier-L with RVL-CDIP pre-training improved the classification accuracy by more than 1.53% over the previous state-of-the-art approach for image-based classification whereas our lightest model DocXClassifier-B presented a 1.25% increase. Additionally, all of our proposed models even performed better than the two-stream combination of EfficientNet and BERT proposed by Ferrando *et al.* (2020) [30]. With only ImageNet pre-training, we achieved an accuracy of 90.14% on the Tobacco3482 dataset, which is not only the highest reported image-based classification accuracy, but also comparable to the recently presented multimodal approach [48] based on the combination of EfficientNet and Hierarchical Attention Networks, which achieved an accuracy of 90.3%.

We also analyzed the class distribution of the DocXClassifier-XL with RVL-CDIP pre-training on the Tobacco3482 dataset, as shown in Fig. 3b. As we can see, in this case there are few classes that are highly misclassified.



	letter	form	email	handwritten	advertisement	scientific report	scientific publication	specification	file folder	news article	budgetv	invoice	presentation	questionnaire	resume	memo
letter	0.93	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.01
form	0.01	0.88	0.00	0.01	0.01	0.02	0.00	0.01	0.00	0.00	0.01	0.02	0.00	0.01	0.00	0.01
email	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
handwritten	0.00	0.01	0.00	0.97	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
advertisement	0.00	0.00	0.00	0.00	0.95	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.01	0.00	0.00
scientific report	0.00	0.02	0.00	0.00	0.00	0.89	0.01	0.01	0.01	0.00	0.01	0.00	0.03	0.00	0.00	0.01
scientific publication	0.00	0.00	0.00	0.00	0.00	0.02	0.95	0.00	0.01	0.01	0.00	0.00	0.01	0.00	0.00	0.00
specification	0.01	0.01	0.00	0.00	0.00	0.01	0.00	0.96	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
file folder	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.97	0.00	0.00	0.00	0.01	0.00	0.00	0.00
news article	0.00	0.00	0.00	0.00	0.02	0.00	0.01	0.00	0.00	0.93	0.00	0.00	0.03	0.00	0.00	0.00
budgetv	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.94	0.01	0.02	0.00	0.00	0.00
invoice	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.95	0.00	0.00	0.00	0.00
presentation	0.00	0.00	0.00	0.00	0.01	0.04	0.00	0.00	0.01	0.01	0.01	0.00	0.89	0.01	0.00	0.00
questionnaire	0.01	0.01	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.92	0.00	0.00
resume	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.98	0.00
memo	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.95

(a) RVL-CDIP

	Letter	Resume	Scientific	ADVE	Email	Report	News	Memo	Form	Note
Letter	0.97	0.00	0.00	0.00	0.00	0.02	0.00	0.01	0.00	0.00
Resume	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Scientific	0.00	0.00	0.83	0.00	0.00	0.06	0.02	0.02	0.06	0.02
ADVE	0.00	0.00	0.00	0.98	0.00	0.02	0.00	0.00	0.00	0.00
Email	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Report	0.08	0.00	0.08	0.00	0.00	0.79	0.00	0.02	0.04	0.00
News	0.00	0.00	0.00	0.03	0.00	0.00	0.97	0.00	0.00	0.00
Memo	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
Form	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.98	0.01
Note	0.07	0.00	0.00	0.00	0.02	0.00	0.00	0.02	0.00	0.88

(b) Tobacco3482

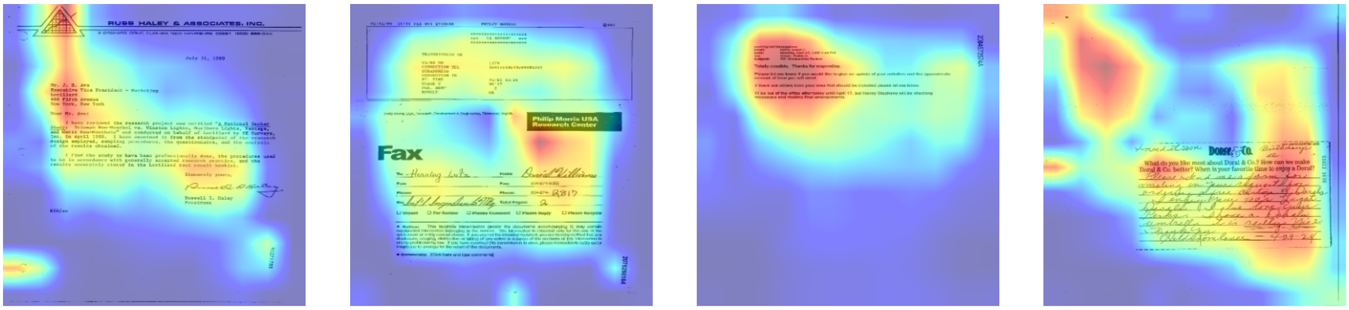
FIGURE 3. The confusion matrices for the DocXClassifier-XL model (with RVL-CDIP pre-training in the case of Tobacco3482) are shown for the two datasets RVL-CDIP and Tobacco3482.

For example, the Scientific class is mainly confused with the Report class, which makes perfect sense since these classes usually have similar visual semantics. This is again very similar to the results of Kanchi *et al.* (2022) [48, Fig. 10] who found a large overlap between the Scientific and Report classes. On the other hand, our approach performs better on the ADVE class than their multimodal approach. This suggests that our visual representations are much richer than the EfficientNet network, since the classification of ADVE class in general depends largely on visual content.

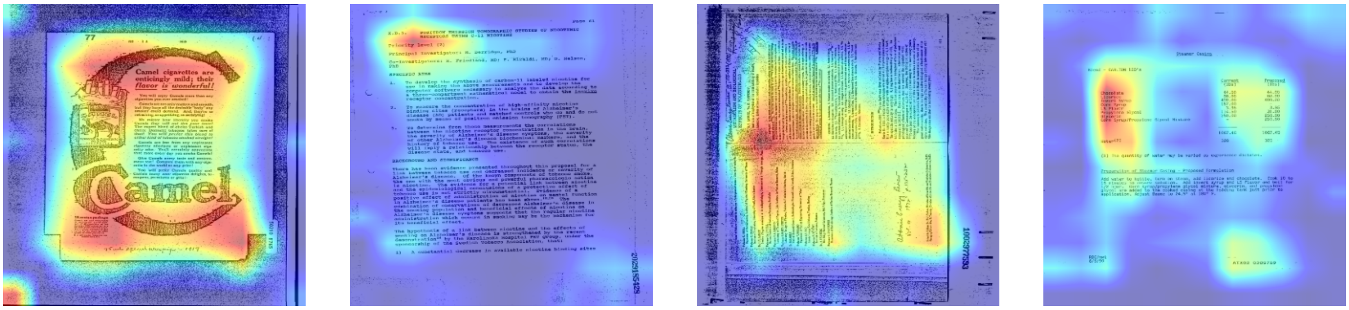
C. ABLATION STUDY

In this section, we present the results of our ablation study, in which we experimented with different sets of configurations to analyze the effects of data augmentation and preprocessing techniques on model performance. The results of the study are summarized in the Table 3. Looking for the best strategy for data augmentation, and training, we started with the base ConvNeXt-B network, a standard input resolution of 224x224, and a simple preprocessing scheme, referred to as $\text{Aug}_{\text{Basic}}$, which involved only downscaling the images to the network resolution, converting the images from grayscale to RGB, and then applying ImageNet normalization. Such a preprocessing scheme has been widely used in the past [10], [11] and therefore provides a good comparison. As can be seen in the table, despite all the modernization, the ConvNeXt model does not perform particularly well with this scheme, achieving only 92.10% accuracy. Adding CutMix and Mixup data during training, denoted by $\text{Aug}_{\text{cutmixup}}$, resulted in a significant increase in network performance from 92.10% to 92.63%. Next, we changed the resolution of the network from 224x224 to 384x384 and trained the network both with and without $\text{Aug}_{\text{cutmixup}}$. It can be seen that increasing the resolution had a very significant effect on performance. The accuracy increased from 92.10% to 93.13% with $\text{Aug}_{\text{basic}}$ and from 92.63% to 93.60% with $\text{Aug}_{\text{cutmixup}}$.

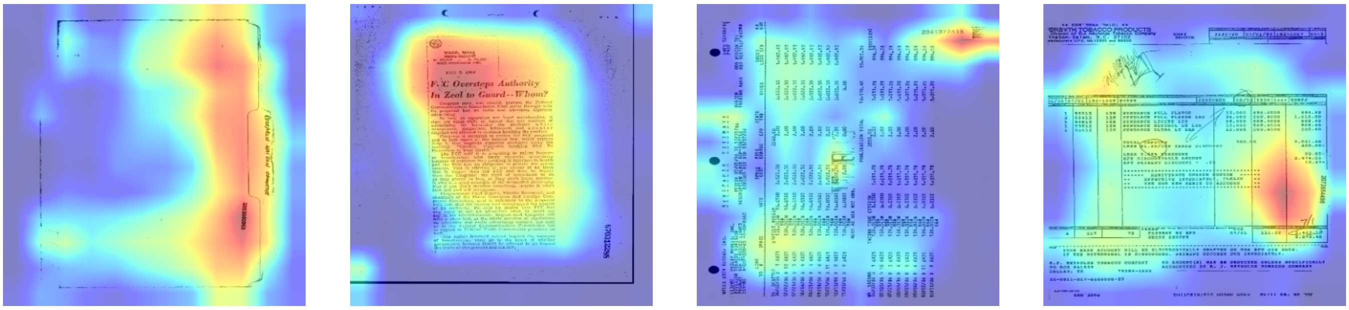
To see how common augmentations applied to natural images affect the classification performance on document images, we replaced $\text{Aug}_{\text{Basic}}$ with a combination of augmentations commonly used to train networks on the ImageNet dataset. We refer to this combination as $\text{Aug}_{\text{ImageNet}}$, which includes RandAugment, and RandomErasing in addition to the basic augmentations. With this replacement, we again trained the network with and without $\text{Aug}_{\text{cutmixup}}$ and report their accuracy. As shown, using $\text{Aug}_{\text{ImageNet}}$ again slightly improves the performance of the network, from 93.13% to 93.21% and from 93.60% to 93.74% with and without $\text{Aug}_{\text{cutmixup}}$ during training, respectively. We then trained the ConvNeXt-L and ConvNeXt-XL networks with this final configuration and report their accuracy. ConvNeXt-L shows no significant improvement over ConvNeXt-B, possibly due to overfitting. As mentioned in Sec. III-C, we computed the accuracy with and without EMA for all three variants. As shown, models with EMA performed significantly better with accuracies of 94.04%, 94.15% and 94.17% than the base models with accuracies of 93.74%, 93.75% and 93.81% respectively. Finally, we replaced the linear classification head of the original ConvNeXt model with a Learned Aggregation Layer to construct the DocXClassifier model and re-trained it as described in Sec. III-C. As can be seen, on RVL-CDIP, we did not notice any change in accuracy when we switched the models from ConvNeXt to DocXClassifier. However, we did find the DocXClassifier model to perform slightly better



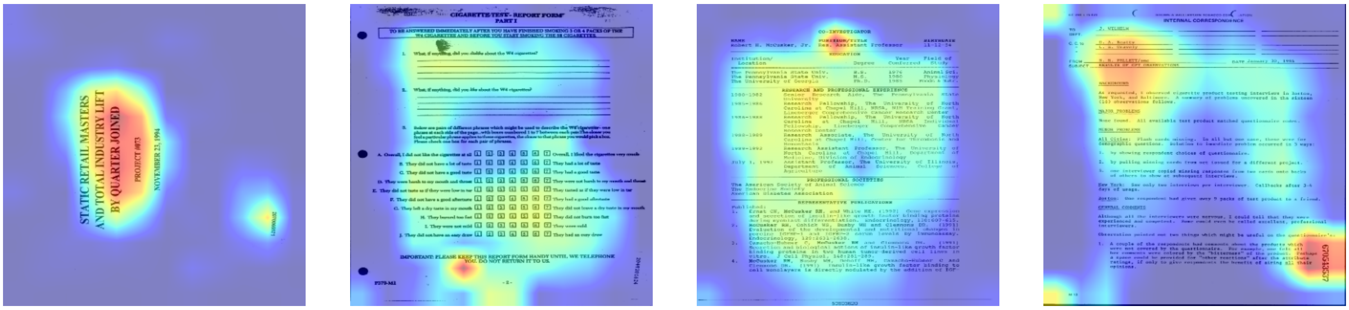
(a) Letter (b) Form (c) Email (d) Handwritten



(e) Advertisement (f) Scientific Report (g) Scientific Publication (h) Specification



(i) File Folder (j) News Article (k) Budget (l) Invoice



(m) Presentation (n) Questionnaire (o) Resume (p) Memorandum

FIGURE 4. Attention maps generated using the DocXClassifier-B model for sample images from each document class from the RVL-CDIP dataset. The intensity of the attention map goes from blue (low) to red (high).

than the base ConvNeXt model on the Tobacco3482 dataset. In addition to performance, we can also see that adding the learned aggregation layer to the models only adds about 8M parameters to the model, which is insignificant compared to the size of the base models. Similarly, when training on Tobacco3482 with ImageNet pre-training, the accuracy was improved slightly from 90.00% to 90.10%.

D. RUNTIME EVALUATION

We also briefly evaluate the runtime performance of our models. In some cases, it may be necessary to use models to classify document images in a real-time scenario, and thus runtime performance can be an important decision factor. We evaluated the runtime performance of our models in terms of throughput on an A100 GPU with a batch size of 256. The results are shown in the Table. 2. As can be seen, the average inference time per image for the DocXClassifier-B, DocXClassifier-L, and DocXClassifier-XL models was 6.5 ms, 10 ms, and 16 ms, respectively. In contrast, the throughput for each model was 153 frames/s, 100 frames/s, and 62 frames/s, respectively. We also see that the inference times of our proposed models are slightly higher than those of previous models such as ResNet-50, VGG-16, and EfficientNet-B4, but we believe this is a small price to pay for the higher performance. Overall, we can conclude that the proposed models are more than suitable for use in real-time scenarios. Note that we did not find any difference in throughput performance between the DocXClassifier models and the corresponding ConvNext models. This makes sense since the number of parameters added by the Learned Aggregation Layer in the DocXclassifier models was rather insignificant compared to the actual sizes of the models.

E. VISUALIZING ATTENTION MAPS

The attention maps generated by our proposed DocXClassifier-B variant for a few samples of the RVL-CDIP dataset are visualized in Fig. 4. It can be observed that the model has learned to focus on specific regions of the image for each class. For example, for the Email, Letter and Memorandum classes, the network has learned to focus on the document header with author, recipients, and subject information. For some classes, network focuses on class-specific information, such as the degrees and experience in Resume class, or blank fields and check-boxes in the Questionnaire class. This shows that our approach is indeed effective in generating human-interpretable attribution maps, eliminating the need to resort to costly post-hoc interpretability approaches to generate the explanations.

V. CONCLUSION

In this work, we have investigated the potential of Transformer-inspired ConvNet designs in combination with advanced data augmentation and training strategies in the context of document image classification. Our study shows that the advanced data augmentation and training techniques commonly used in natural image classification can be directly applied to document image classification and lead to significant performance improvements. Moreover, our work using only visual features outperforms several existing multimodal approaches. This suggests that these multimodal techniques either do not exploit the full potential of the multimodality of the data, or that their feature generation backbones (visual and textual) still need independent improvement in the context of document classification. In contrast to previous ConvNet-based approaches, we have also introduced in our models the ability to generate class-specific attention maps, which makes them inherently interpretable and opens a new avenue for explainable classification of document images. It is worth mentioning that our proposed approach can also be adapted to other application domains with slight domain-specific modifications, directly benefiting other domains where the inherent interpretability of ConvNets is desired. A plausible future direction of our work could be to use our visual backbones in combination with existing textual backbones to see if they can improve overall classification performance in a multimodal setting. Another future direction could be to improve the model design so that finer attention maps can be created without sacrificing performance.

REFERENCES

[1] F. Cesarini, M. Lastri, S. Marinai, and G. Soda, “Encoding of modified x-y trees for document classification.,” pp. 1131–1136, 01 2001.

[2] J. van Beusekom, D. Keysers, F. Shafait, and T. M. Breuel, “Distance measures for layout-based document image retrieval,” Second International Conference on Document Image Analysis for Libraries (DIAL’06), pp. 11 pp.–242, 2006.

[3] S. Marinai, Introduction to Document Analysis and Recognition, pp. 1–20. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.

[4] A. Dengel and F. Dubiel, “Clustering and classification of document structure-a machine learning approach,” Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, vol. 2, pp. 587–591, 1995.

[5] S. Chen, Y. He, J. Sun, and S. Naoi, “Structured document classification by matching local salient features,” Proc. - Int. Conf. Pattern Recognit., no. Icpr, pp. 653–656, 2012.

[6] J. Kumar, P. Ye, and D. Doermann, “Learning document structure for retrieval and classification,” Proc. - Int. Conf. Pattern Recognit., pp. 1558–1561, 2012.

[7] S. Baldi, S. Marinai, and G. Soda, “Using tree-grammars for training set expansion in page classification,” Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, vol. 2003-Janua, no. Icdar, pp. 829–833, 2003.

[8] M. Diligenti, P. Frasconi, and M. Gori, “Hidden tree Markov models for document image classification,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 25, no. 4, pp. 519–523, 2003.

[9] L. Kang, J. Kumar, P. Ye, Y. Li, and D. Doermann, “Convolutional neural networks for document image classification,” Proc. - Int. Conf. Pattern Recognit., pp. 3168–3172, 2014.

[10] 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,” Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, vol. 2015-Novem, pp. 1111–1115, 2015.

[11] A. W. Harley, A. Ufkes, and K. G. Derpanis, “Evaluation of deep convolutional nets for document image classification and retrieval,” Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, vol. 2015-Novem, pp. 991–995, 2015.

[12] M. N. Asim, M. U. G. Khan, M. I. Malik, K. Razzaque, A. Dengel, and S. Ahmed, “Two stream deep network for document image classification,” Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, pp. 1410–1416, 2019.

[13] N. Audebert, C. Herold, K. Slimani, and C. Vidal, “Multimodal deep networks for text and image-based document classification,” in Commun. Comput. Inf. Sci., vol. 1167 CCIS, pp. 427–443, Springer, Cham, sep 2020.

[14] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, “LayoutLM: Pre-training of Text and Layout for Document Image Understanding,” Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., vol. 20, pp. 1192–1200, 2020.

[15] Y. Xu, Y. Xu, T. Lv, L. Cui, F. Wei, G. Wang, Y. Lu, D. Florencio, C. Zhang, W. Che, M. Zhang, and L. Zhou, “LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding,” pp. 2579–2591, Association for Computational Linguistics (ACL), dec 2021.

[16] R. Powalski, Ł. Borchmann, D. Jurkiewicz, T. Dwojak, M. Pietruszka, and G. Pałka, “Going Full-TILT Boogie on Document Understanding with

Text-Image-Layout Transformer," in Doc. Anal. Recognit. – ICDAR 2021, vol. 12822 LNCS, pp. 732–747, 2021.
[17] S. Bakkali, Z. Ming, M. Coustaty, and M. Rusinol, "Visual and textual deep feature fusion for document image classification," pp. 2394–2403, 06 2020.
[18] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," 2022.
[19] H. Touvron, M. Cord, A. El-Nouby, P. Bojanowski, A. Joulin, G. Synnaeve, and H. Jégou, "Augmenting convolutional networks with attention-based aggregation," 2021.
[20] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in International Conference on Learning Representations, 2018.
[21] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "Cutmix: Regularization strategy to train strong classifiers with localizable features," pp. 6022–6031, 10 2019.
[22] E. Cubuk, B. Zoph, J. Shlens, and Q. Le, "Randaugment: Practical automated data augmentation with a reduced search space," pp. 3008–3017, 06 2020.
[23] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "Deep networks with stochastic depth," in Computer Vision – ECCV 2016 (B. Leibe, J. Matas, N. Sebe, and M. Welling, eds.), (Cham), pp. 646–661, Springer International Publishing, 2016.
[24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," 06 2016.
[25] D. Shin, Christian and Doermann, "Document Image Retrieval Based on Layout Structural Similarity.," Proc. 2006 Int. Conf. Image Process. Comput. Vision, Pattern Recognit., vol. 2, pp. 606–612, 2016.
[26] J. Kumar, P. Ye, and D. Doermann, "Structural similarity for document image classification and retrieval," Pattern Recognit. Lett., vol. 43, no. 1, pp. 119–126, 2014.
[27] N. Chen and D. Blostein, "A survey of document image classification: Problem statement, classifier architecture and performance evaluation," Int. J. Doc. Anal. Recognit., vol. 10, no. 1, pp. 1–16, 2007.
[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255, 2009.
[29] M. Z. Afzal, A. Kolsch, S. Ahmed, and M. Liwicki, "Cutting the Error by Half: Investigation of Very Deep CNN and Advanced Training Strategies for Document Image Classification," Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, vol. 1, pp. 883–888, 2017.
[30] J. Ferrando, J. L. Domínguez, J. Torres, R. García, D. García, D. Garrido, J. Cortada, and M. Valero, "Improving accuracy and speeding up document image classification through parallel systems," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12138 LNCS, pp. 387–400, 2020.
[31] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proceedings of the 36th International Conference on Machine Learning (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of Proceedings of Machine Learning Research, pp. 6105–6114, PMLR, 09–15 Jun 2019.
[32] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.
[33] S. A. Siddiqui, A. Dengel, and S. Ahmed, "Analyzing the potential of zero-shot recognition for document image classification," in Document Analysis and Recognition – ICDAR 2021 (J. Lladós, D. Lopresti, and S. Uchida, eds.), (Cham), pp. 293–304, Springer International Publishing, 2021.
[34] T. Dauphinee, N. Patel, and M. Rashidi, "Modular multimodal architecture for document classification," 2019.
[35] Y. Xiong, Z. Dai, Y. Liu, and X. Ding, "Document image classification method based on graph convolutional network," in Neural Information Processing (T. Mantoro, M. Lee, M. A. Ayu, K. W. Wong, and A. N. Hidayanto, eds.), (Cham), pp. 317–329, Springer International Publishing, 2021.
[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems (F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), vol. 25, Curran Associates, Inc., 2012.
[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in International Conference on Learning Representations, 2015.
[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.
[39] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," 2021.
[40] I. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, M. Lucic, and A. Dosovitskiy, "Mlp-mixer: An all-mlp architecture for vision," 2021.
[41] C. Tang, Y. Zhao, G. Wang, C. Luo, W. Xie, and W. Zeng, "Sparse mlp for image recognition: Is self-attention really necessary?," 2021.
[42] S. Fekri-Ershad, "Bark texture classification using improved local ternary patterns and multilayer neural network," Expert Systems with Applications, vol. 158, p. 113509, 2020.
[43] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation policies from data," 2019.
[44] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," 2017.
[45] 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," Proc. - Int. Conf. Pattern Recognit., vol. 2018-Augus, pp. 3180–3185, 2018.
[46] R. Sarkhel and A. Nandi, "Deterministic routing between layout abstractions for multi-scale classification of visually rich documents," pp. 3360–3366, 08 2019.
[47] T. Dauphinee, N. Patel, and M. Rashidi, "Modular multimodal architecture for document classification," 12 2019.
[48] S. Kanchi, A. Pagani, H. Mokayed, M. Liwicki, D. Stricker, and M. Z. Afzal, "Emmdocclassifier: Efficient multimodal document image classifier for scarce data," Applied Sciences, vol. 12, p. 1457, 01 2022.
[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.
[50] C. Tensmeyer and T. Martinez, "Analysis of Convolutional Neural Networks for Document Image Classification," Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, vol. 1, pp. 388–393, 2017.
[51] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Going deeper with image transformers," 2021.

SAIFULLAH received the B.S. degree in mechanical engineering and the M.S. degree in robotics and intelligent machine engineering from the National University of Sciences and Technology (NUST), Pakistan. He is currently pursuing his Ph.D. at the University of Kaiserslautern and is working as a researcher at the German Research Center for Artificial Intelligence (DFKI GmbH) under the supervision of Prof. Dr. Prof. H. C. Andreas Dengel. His research interests include document understanding and analysis, explainability and robustness of deep learning models, and privacy preservation in deep learning.

STEFAN AGNE received the Diploma degree in computer science and the Ph.D. degree under the supervision of Dr. A. Dengel from the University of Kaiserslautern, Germany, in 1995, and the Ph.D. degree from the University of Bern, in 2008, under the supervision of Dr. H. Bunke. Since 1992, he has been working in document analysis and understanding with the German Research Center for Artificial Intelligence (DFKI GmbH), Germany. He is currently leading the topic area pattern recognition with the Smart Data and Knowledge Services Department, German Research Center for Artificial Intelligence, Kaiserslautern.

ANDREAS DENGEL received the Diploma degree in C.S. from the University of Kaiserslautern and the Ph.D. degree from the University of Stuttgart. He is the Scientific Director at the German Research Center for Artificial Intelligence (DFKI GmbH), Kaiserslautern. In 1993, he became a Professor at the Computer Science Department, University of Kaiserslautern, where he is currently the Chair of Knowledge-Based Systems, and since 2009, he has been a Professor (Kyakuin) at the Department of Computer Science and Information Systems, Osaka Prefecture University. He also worked at IBM, Siemens, and Xerox Parc. He is an author of more than 300 peer-reviewed scientific publications and supervised more than 170 Ph.D. and master's theses. His main scientific emphasis is in the areas of pattern recognition, document understanding, information retrieval, multimedia mining, semantic technologies, and social media. He is a member of several international advisory boards, chaired major international conferences, and founded several successful start-up companies. He is an IAPR Fellow and received prominent international awards. Moreover, he is a co-editor of international computer science journals and has written or edited 12 books.

SHERAZ AHMED is Senior Researcher at DFKI GmbH in Kaiserslautern, where he is leading the area of Time Series Analysis and Life Science. He received his MS and PhD degrees in Computer Science from TUK, Germany under the supervision of Prof. Dr. Prof. h.c. Andreas Dengel and Prof. Dr. habil. Marcus Liwicki. His PhD topic is Generic Methods for Information Segmentation in Document Images. Over the last few years, he has primarily worked on development of various systems for information segmentation in document images. His research interests include document understanding, generic segmentation framework for documents, pattern recognition, anomaly detection, Gene analysis, medical image analysis, and natural language processing. He has more than 100 publications on the said and related topics including three journal papers and two book chapters. He is a frequent reviewer of various journals and conferences including Pattern Recognition Letters, Neural Computing and Applications, IJDAR, ICDAR, ICFHR, and DAS.

• • •