SN Computer Science (2023) 4:5
https://doi.org/10.1007/s42979-022-01414-4

ORIGINAL RESEARCH

Document Image Analysis Using Deep Multi-modular Features

K. V. Jobin[1] · Ajoy Mondal[1] · C. V. Jawahar[1]

Received: 13 April 2022 / Accepted: 15 September 2022

Abstract
Texture or repeating patterns, discriminative patches, and shapes are the salient features for various document image analysis problems. This article proposes a deep network architecture that independently learns texture patterns, discriminative patches, and shapes to solve various document image analysis tasks. The considered tasks are document image classification, genre identification from book covers, scientific document figure classification, and script identification. The presented network learns global, texture, and discriminative features and combines them judicially based on the nature of the problems to be solved. We compare the performance of the proposed approach with state-of-the-art techniques on multiple publicly available datasets such as Book-Cover, RVL-CDIP, CVSI and DOCFigure. Experiments show that our approach outperforms state-of-the-art for the genre and document figure classifications and obtains comparable results for document image and script classification tasks.

Keywords Document analysis · Texture feature · Document image classification · Script identification · Document figure classification · Identification of the book's genre

Introduction

Features play an essential role in various classification tasks related to document image analysis [1–4], such as document image classification [1, 2, 5–7], document figure classification [3], book cover classification [8], and script identification [4, 9, 10]. Researchers have proposed various approaches to tackle these classification tasks [11]. Each approach focuses on a specific problem, thereby fails to take cognizance of other problems. A Feature responsible for one particular task may not perform well on other tasks.

In script identification problems, characters or unique curves in a language are the key features for distinguishing it from other languages. Figure 1c shows sample images of various languages. The texture features uniquely represent the curves and characters of a script to solve this problem [4, 9, 10]. On the other hand, most of the existing approaches to document image classification tasks focus on global shape features [1, 12–14] that capture the arrangement of various logical regions, such as heading, paragraph, and figure. For example, the shape of documents with double-column formatting differs from documents with single-column formatting. Hence, shape features are more useful than texture features for solving this problem. However, both the texture and shape features are required to solve document figure classification [3] and book cover classification [8] tasks.

It is well established that the features corresponding to local image regions (patches) are more discriminative than the global features for various tasks [15]. These patches are called discriminative patches. Researchers designed various deep networks to extract discriminative features for fine-grained image classification tasks [15, 16]. These architectures extract only a specific features (global, texture, or local discriminative) to solve a particular type of problem. Hence, the existing methods lack a generic model that extracts all the features discussed to solve the aforesaid tasks.

This work proposes a general-purpose deep network architecture to extract three types of features (i.e., global, texture, and local). We consider four different document image analysis tasks such as (1) document image

✉ K. V. Jobin
jobin.kv@research.iiit.ac.in

Ajoy Mondal
ajoy.mondal@iiit.ac.in

C. V. Jawahar
jawahar@iiit.ac.in

[1] Center for Visual Information Technology, International Institute Information Technology, Hyderabad, Telangana 500032, India

Published online: 15 October 2022

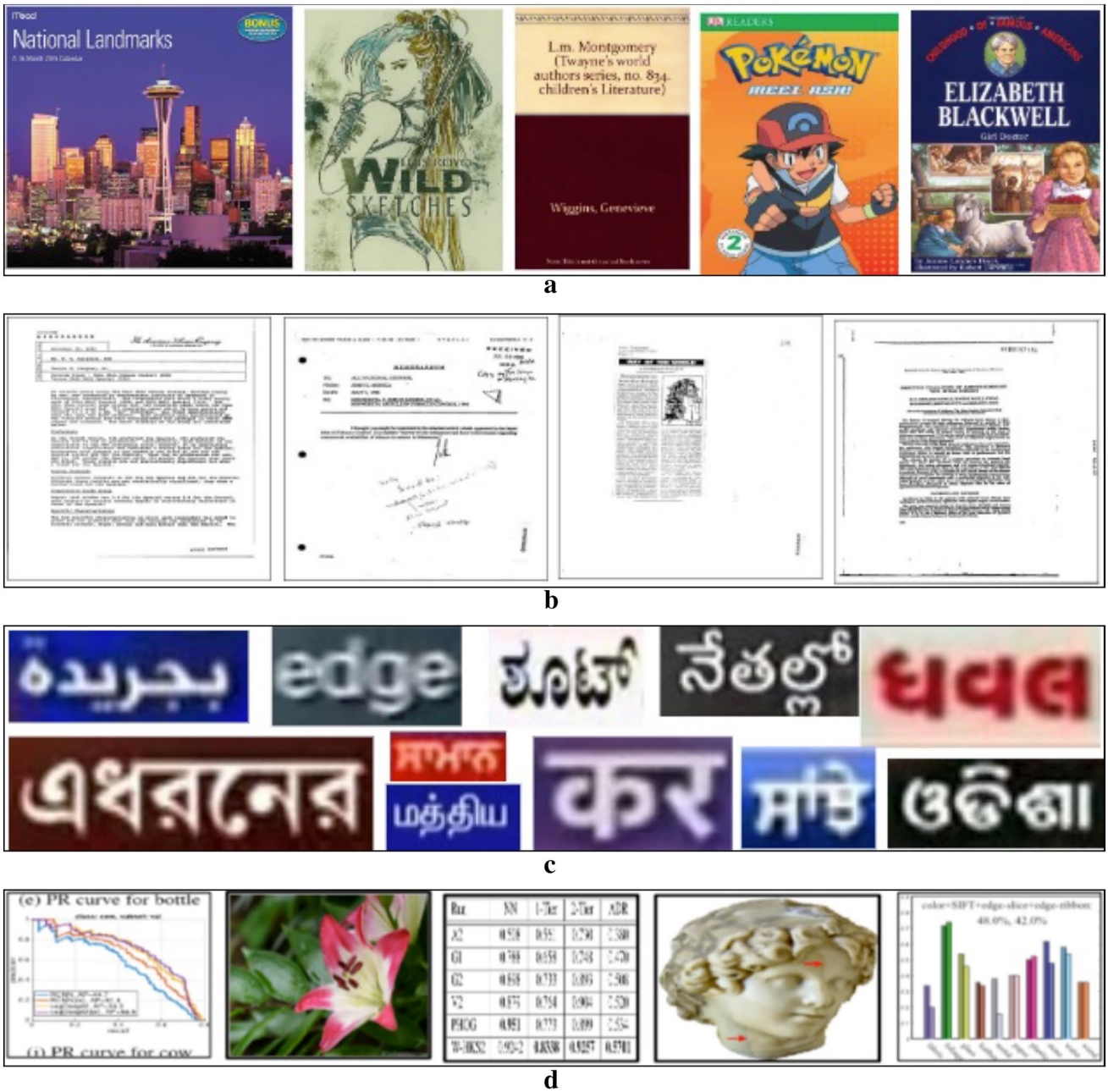

Fig. 1 Shows sample images from various considered datasets—**a** Book-Cover, **b** RVL-CDIP, **c** CVSI, and **d** DOCFigure

classification, (2) genre classification, (3) scientific document figure classification, and (4) script identification. We choose these four tasks to show the effectiveness of our deep multi-modular feature on document image analysis, mainly the classification of the document images at various units. The document image and genre classifications act on whole document images. The document figures (e.g., Bar chart, Natural image, Pie chart, etc.) and words are smaller document units. The proposed approach performs various classification tasks by a judicial combination of these features. The extensive experiments on public benchmark datasets show that the proposed method outperforms the state-of-the-art on genre and document figure classification tasks and obtains comparable results on document image and script classification tasks.

Related Work

Document Image Classification

The researchers solve the document image classification tasks using various types of features—(1) visual feature, (2) textual feature, (3) combination of textual and spatial features, and (4) combination of visual, textual, and spatial features. The approaches used only visual features and no external feature extraction modules such as OCR and layout detector. Harley et al. [12] solve document image classification tasks using deep Convolution Neural Networks (CNNS). The researchers improve the performance of the document image classification task using various CNN architectures such as VGG-16 [2], AlexNet [13], GoogLeNet [14], InceptionResNetV2 [17], and LadderNet [18]. There is a limitation to improving accuracy by changing network architectures. Das et at. [1] use a stacked generalized ensemble technique to combine predictions generated by different base networks.

Other than visual features, several methods explore textual and spatial features for classifying document images. Xu et al. [19] successfully utilize the language representation model (BERT) [20] for classifying document images. This model interacts between the textual and layout features to improve classification accuracy. Works [19, 21, 22] combine visual, textual, and spatial features for document image classification and obtain the best results. All the above-discussed methods utilize only global visual features and not local discriminative and texture features for document image classification. Instead of global visual features, sometimes local discriminative and texture features help to discriminate one category of documents from other categories. Our work aims to extract three visual features and judicially combine them for the classification task.

Genre Classification from Book Cover

In recent years, there has been an increasing interest in automated genre classification based on images by leveraging the strength of the deep neural network. Iwana et al. [8] introduce a Book-Cover dataset with 30 genres and solve it using the AlexNet [23] pre-trained on ImageNet [23]. Zujovic et al. [24] classify paintings based on genres. The authors utilize gray level features and color features from the images and fed these features to different classifiers for prediction. Holly et al. [25] use textual features along with visual features in a transfer learning framework to classify the genre of the book covers. In the same direction, Biradar et al. [26] use both the textual and image features to determine the book's genre. The work [27] benchmarks a set of state-of-the-art image classification models for book cover classification.

Document Figure Classification

Figure classification is the primary step for information retrieval/extraction from a figure in a document image. Various figures like charts, tables, and natural images are used to visually represent a wide range of textual information in books, scientific articles, newspapers, etc. Text recognition using Optical Character Recognition (OCR) is the primary process for understanding the content of the document images. Increasing use of figures in documents suggests figure classification can be an important sub-task for OCR for a better and complete understanding of the document images [28]. In early works [29–33], different handcrafted features are used to recognize various types of charts in the document images.

Zhou et al. [29, 30] considered Hough transformation to recognize bar charts in the document images. Prasad et al. [32] considered SIFT and HOG features to recognize five different types of chart images. Due to the large visual similarity among subordinate categories, the handcrafted features fail to achieve good accuracy on the figure classification task.

To solve the limitation of handcrafted features for the figure classification task, recently, Kavasidis et al. [34] proposed a saliency-based Convolutional Neural Network (CNN) for localizing different types of figures in the documents. This work is limited to localize tables, Bar charts, and Pie charts. Tang et al. [35] proposed a novel framework (DeepChart) to classify charts by combining (CNNS) and Deep Belief Networks (DBNS). The authors experimentally established that their method is far better than handcrafted features. In the same direction, Siegel et al. [36] proposed various document figure classification algorithms using the learnable features. Similarly, the work [37] used a deep neural network to rank the figures. The work [38] focuses on the full-text indexing of all text referring to the images and filtering for disciplines and image type.

Script Identification

Script identification is an inevitable step for text understanding under multi-lingual scenarios. The main challenge for script identification is the interference caused by the similarity between texts in different languages. Other are complex background, noise, and low resolution. Pre-deep learning approaches used handcrafted features to identify the script. Shijian et al. [39] proposed a document vectorization technique that transformed documents into electronic document

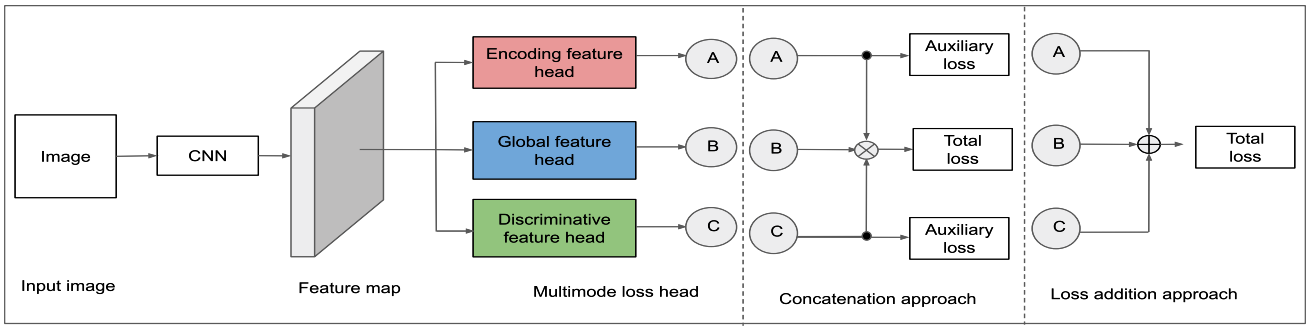

Fig. 2 Presents a block diagram of the proposed approach. The input image passes through the layers of convolutional filters of an CNN architecture to extract convolutional features. From the convolutional feature, the model extracts three different modalities of features: an encoding feature, a global feature, and a discriminative feature. Here, ⊗ and ⊕ indicate concatenation and loss addition of features, and • represents sharing of the same features, respectively

vectors. Therefore, scripts/languages are identified using vertical component and shape analysis. The properties of the connected component are used to identify the script in [40]. Similarly, Sharma et al. [41] extracted local binary pattern, histogram of oriented gradient, and gradient local auto-correlation features and exploits SVMs and ANNs to classify the scripts.

Advancement in deep neural architectures motivates researchers to use network for script identification. Mei et al. [42] integrated CNN and LSTM-RNN to identify the scripts of natural scene images. Lu et al. [43] discussed a method for script identification by integrating the local and the global CNN features. The local features help to extract subtle differences from the scripts. The final decision was obtained by combining the results using the AdaBoost algorithm. Shi et al. [10] introduced a discriminative convolution network that can identify subtle discriminative feature for the script. Bhunia et al. [44] used attention-based Convolutional-LSTM network for script identification. Ghosh et al. [45] proposed lightweight script identification network to identify video scripts.

Deep Multi-modular Features

The traditional deep convolutional neural networks (CNNs) such as VGG-M [46], VGG-V [47], ResNet [48] and GoogleNet [49] learned the generic global features from the given input images. The learned global features may not always be capable of efficiently representing the input image's textural nature and enhancing the discriminative power of the networks [50]. As a result, these generic global features fail to classify sub-categories presented in an object category.

We propose a deep network shown in Fig. 2 capable of extracting global, discriminative local, textural features for document image classification. The proposed model consists of two blocks—(1) the first block consists of several convolution layers, and (2) the second block consists of three different heads, mainly the global feature head, discriminative feature head, and encoding feature head. Each of the heads extracts a specific type of feature. The global feature head extracts the global feature. The discriminative feature head extracts the features corresponding to the discriminative local image patches. While the encoding feature head learns the encoded feature. Finally, the three types of extracted features are concatenated to represent the discriminative feature.

Encoding Feature Head

The texture feature extraction using deep learning proposed by Cimpoi et al. [50] lacks end-to-end training. Since, the model is not end-to-end trainable, it may not learn the discriminative texture features. To get discriminative texture features, we adapt the encoding network called Deep Texture Encoding Network (Deep TEN), proposed by Zhang et al. [51]. This network learns a codebook $C = \{c_1, c_2, c_3, \ldots, c_K\}$ containing K codewords, where $c_i \in \mathbb{R}^D$ and smoothing factor $s = \{s_1, s_2, s_3, \ldots, s_K\}$ from a set of feature vector $X = \{x_1, x_2, x_3, \ldots, x_N\}$, where $x_i \in \mathbb{R}^D$. Irrespective of the number of feature vectors N, the encoder output is a fixed length representation $E = \{e_1, e_2, e_3, \ldots, e_K\}$, where $e_i \in \mathbb{R}^D$, and each e_k is calculated using Eq. (1):

$$e_k = \sum_{i=1}^{N} e_{ik} = \sum_{i=1}^{N} a_{ik} r_{ik}, \tag{1}$$

where r_{ik} is calculated by $r_{ik} = x_i - c_k$ and a_{ik} is the weight associated with each pair of x_i and c_k and it is calculated by Eq. (2):

$$a_{ik} = \frac{exp(-s_k||r_{ik}||^2)}{\sum_{j=1}^{K} exp(-s_j||r_{ij}||^2)}. \tag{2}$$

<table>
<tr><th colspan="2">Discriminating head</th><th>Global head</th><th>Encoding head</th></tr>
<tr><td colspan="2"></td><td>7×7, 64, stride 2</td><td rowspan="5"></td></tr>
<tr><td colspan="2"></td><td>3×3 max pool, stride 2</td></tr>
<tr><td colspan="2"></td><td>$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128, C{=}32 \\ 1\times1, 256 \end{bmatrix} \times3$</td></tr>
<tr><td colspan="2"></td><td>$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256, C{=}32 \\ 1\times1, 512 \end{bmatrix} \times4$</td></tr>
<tr><td colspan="2"></td><td>$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512, C{=}32 \\ 1\times1, 1024 \end{bmatrix} \times6$</td></tr>
<tr><td></td><td>$1\times1, m\times L$</td><td rowspan="2">$\begin{bmatrix} 1\times1, 1024 \\ 3\times3, 1024, C{=}32 \\ 1\times1, 2048 \end{bmatrix} \times3$</td><td>$1\times1, 512$</td></tr>
<tr><td>CCP layer
$1\times1, L$</td><td>max pool
$1\times1, L$</td><td>Encoding layer
$L-d$ fc</td></tr>
<tr><td colspan="2"></td><td colspan="2">global average pool, $L-d$ fc
$4L-d$ fc , softmax</td></tr>
</table>

Fig. 3 Presents the detailed architecture of the proposed network. The thicker horizontal line represents a branching point where all the blocks touching a thicker horizontal line are connected

The output of the encoder E is differentiable *w.r.t* the input X, codebook C and smoothing factor s. The codebook C and smoothing factor s are initialized with the random values within the range $(\frac{-1}{\sqrt{K\times D}}, \frac{1}{\sqrt{K\times D}})$ and $(\frac{-1}{\sqrt{K}}, \frac{1}{\sqrt{K}})$, respectively. The encoder is differentiable *w.r.t* the loss function and learn the codebook and smoothing factor in a supervised manner.

Discriminative Feature Head

We propose a discriminative feature head to learn features corresponding to the discriminative image patches, inspired by the work [15]. The discriminative feature head consists of an asymmetric two-stream architecture. The discriminative patch learning is a 1×1 convolutional layer followed by a Global Max Pooling (GMP) layer and a classification layer (fully connected layer and a softmax layer). This architecture is not guaranteed to fire at discriminative patches as desired. A Cross-Channel Pooling (CCP) layer followed by a softmax layer is introduced to learn class-specific discriminative image patches.

The CCP layer is an average pooling layer. Instead of a fully connected classification layer, each classification output node's value is calculated by averaging the consecutive m values of the output of GMP layer. Hence, the number of output channels of the GMP layer should be $m\times L$, where L is the number of classes, and m is the number of filters per class. Since there are no learnable parameters in the cross-channel pooling layer, the 1×1 convolutional layer weights are adjusted directly via both loss functions of the classification layer and the softmax layer.

Global Feature Head and Backbone Neural Network

The global feature head consists of a convolutional layer, a fully connected layer, and a softmax layer. It captures the structural attributes and global nature of the input image. It is the same as the traditional image classification networks such as VGG-M [46], VGG-V [47], ResNet [48] and GoogleNet [49]. Like ResNet [48], we use a global average pooling layer before the fully connected layer. Hence, the proposed architecture can handle arbitrarily shaped images. Figure 3 shows the detail of the proposed architecture.

Network Design

We have two design choices to combine the encoding, discriminative, and global feature heads. These are—(1) adding the losses calculated with individual feature heads as suggested in [15] and (2) concatenating all the three feature

heads followed by a linear layer to calculate the loss. In the first choice, the total loss (L_{total}) can be calculated using Eq. (3)

$$L_{\text{total}} = L_{\text{enc}} + L_{\text{dis}} + L_{\text{glob}}. \tag{3}$$

The direct sum of all the losses is a way to combine the various feature head. The contribution of multiple feature head losses depends on the numerous problems. For example, the encoding feature head loss is more relevant than the global feature head loss for a script identification in a word image. The second option of concatenating all the three feature heads, followed by a linear layer to calculate the loss. In this option, the linear layer learns the weight for each feature head based on the problem. The discriminative feature head fails to understand the discriminative patches for two reasons. These are—(1) the discriminative feature head consists of asymmetric two-stream architecture, and (2) direct loss from the label is required as explained in Sect. 3.2 to learn the CCP layer.

To overcome these issues, we propose an auxiliary loss to learn the discriminative feature head as introduced in [52]. Similarly, we train the encoding feature head with an auxiliary loss to calculate between the labels and the prediction of the encoding head. Apart from the global feature head using softmax loss, another classifier is applied to learn the encoding and discriminative feature heads individually as an auxiliary loss. The additional loss helps optimize the learning process, while the master branch loss takes the most responsibility. Figure 2 illustrates the detailed architecture of the proposed model.

Training Details

We train the proposed network using Stochastic Gradient Descent (SGD) with a momentum of 0.9 and a learning rate updated based on Cosine annealing strategy proposed in [53]. The initial learning rate η is set to 0.001 and after each epoch the learning rate is updated based on Eq. (4)

$$\eta_t = \frac{1}{2}\left(1 + \cos\left(\frac{t\pi}{T}\right)\right)\eta. \tag{4}$$

The number of codewords, K in the encoder layer is set to 64 and the number of channels, m per class in the Cross-Channel pooling layer is set to 20. The size is set to 32 and the total number of iterations set to $100K$.

Experiments

We run all the experiments in an Intel i7 CPU processor with Nvidia GTX 1080Ti GPU, and our program utilizes 3259MiB of GPU memory and 2240MiB of CPU memory for the batch size of 32 during the training. We evaluate the performance of the proposed network architecture for four different tasks—(1) document image classification, (2) genre classification, (3) script identification, and (4) scientific document figure classification.

Document Image Classification

In this task, our goal is to assign a pre-defined category label (like Advertisement, Email, Form, Letter, and Memo) to a given document image. It is often a prerequisite step towards high-level document image analysis tasks. Various types of document images contain distinct textural properties. More specifically, the different categories of document images can be discriminated against concerning the features corresponding to their local patches. This experiment combines global, discriminative local, and texture features to classify a given document into a pre-defined category.

Dataset and Pre-processing

We use the existing benchmark RVL-CDIP[1] [12] dataset to analyze the performance of the proposed approach on document image classification task. The dataset consists of scanned grayscale images of 16 categories of documents from lawsuits against American Tobacco companies. The dataset is divided into training, validation, and test sets, each containing 320K, 40K, and 40K images. The sample images of this dataset are shown in Fig. 1b.

As discussed in [13], we apply similar pre-processing steps to the dataset. The steps are—(1) the images are resized to 384×384, and (2) we duplicate a single image channel into three channels for network compatibility. Figure 4 shows the challenges of intra-class dissimilarity and inter-class similarity present in the RVL-CDIP dataset. The first row of Fig. 4 highlights that different categories of documents (e.g., Advertisement, News article, Presentation, Scientific report, and Specification) have similar structural properties. While the second row of Fig. 4 highlights that documents of one particular category (e.g., Questionnaire) have distinct structural properties. High structural similarity among various classes of documents and high structural dissimilarity among document images of a specific category present in RVL-CDIP dataset makes document image classification tasks more complicated.

Ablation Study

Table 1 shows the ablation study to understand the contributions of various feature heads in document image classification tasks. The table also shows the improved performance

[1] http://www.cs.cmu.edu/~aharley/rvl- cdip/.

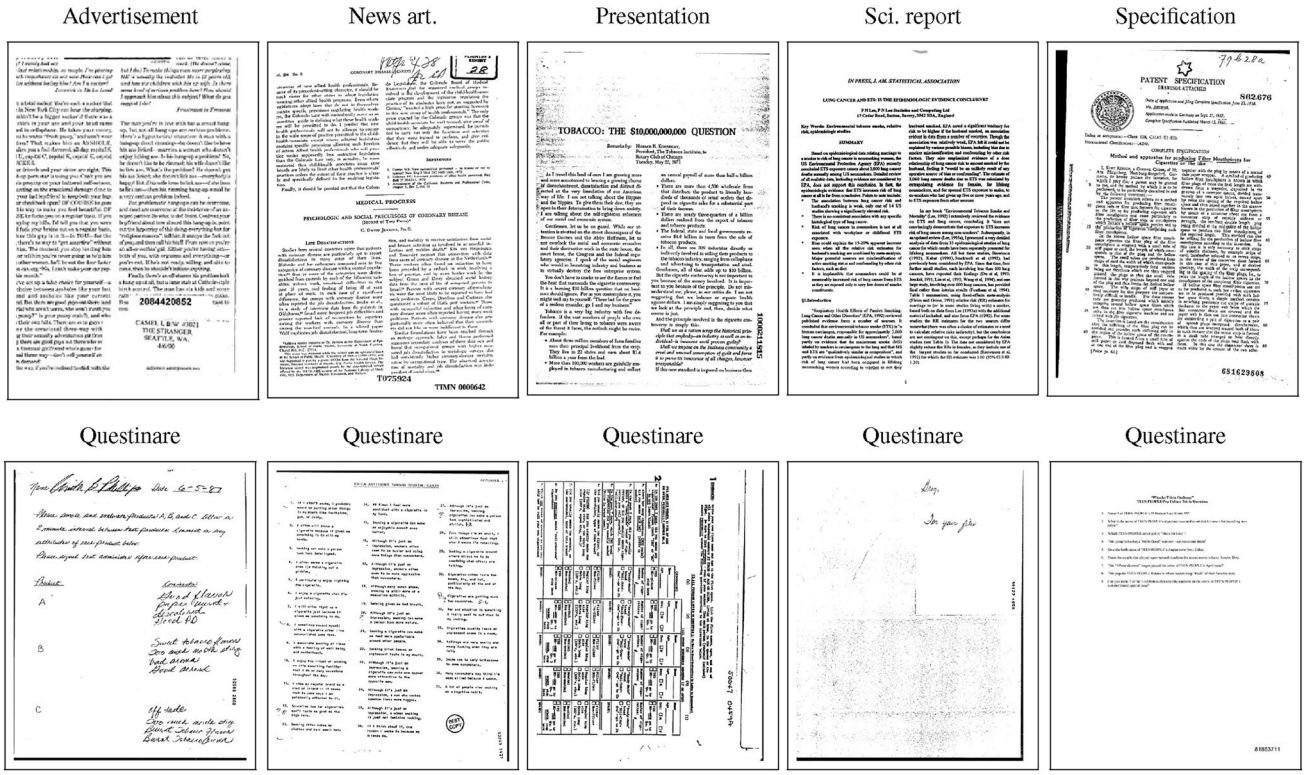

Fig. 4 Shows challenges on the RVL-CDIP dataset. The first row indicates document images from different categories share similar structural properties. The second row highlights the document images from a specific type, namely "Questionnaire" with the separate structural property

Table 1 Shows the document classification accuracy under various settings, where ⊕ represents the addition of the losses separately, and ⊗ represents the concatenation of all in the last layers followed by a linear neural network

Feature	Mode	Accuracy↑
Global		91.18%
Global + discriminative	⊗	92.23%
Global + encoding	⊗	91.21%
Global + encoding + Discriminative	⊗	92.67%
Global + discriminative	⊕	92.44%
Global + encoding	⊕	91.83%
Global + encoding + discriminative	⊕	**92.94%**

Bold indicates best result

of adding separate feature heads over the feature head's concatenation. The base network provides an accuracy of 91.18%. The concatenation of discriminating feature head to the global loss improved the accuracy by 1.05%. Adding the discriminative head to the global head improved the accuracy by 1.26%. The table also highlights that the addition of a discriminative feature head to the global feature head in both cases improves the accuracy (i.e., 92.23% and 92.44%)as compared to adding an encoding head to the global feature head (i.e., 91.21% and 91.83%). This result shows that the discriminating feature head is more informative than the encoding and texture features for the document image classification problem. However, the three feature heads' combination improves performance further in both cases (92.67% and 92.94%).

Learned Feature Visualization

We visualize the effects of the discriminative feature head shown in Fig. 3 for solving document image classification. For this purpose, we calculate the L_2 norm of the 1×1, $m \times L$ convolutional layer (shown in Fig. 3). We create a $2D$ heat map using the L_2 norm values to visualize the discriminative patches learned by the discriminative head. Similarly, we also visualize the L_2 norm calculated at the feature obtained after the 1×1, 1024 convolutional layer in the global feature head (shown in Fig. 3). Figure 5 shows the heat map, which is overlaid on the input image to visualize the learned discriminative patches and the global features of the input image. From Fig. 5, we observe that the discriminative feature head concentrates on the meaningful discriminating regions of the image, which helps the network to differentiate one particular category of the document from other classes.

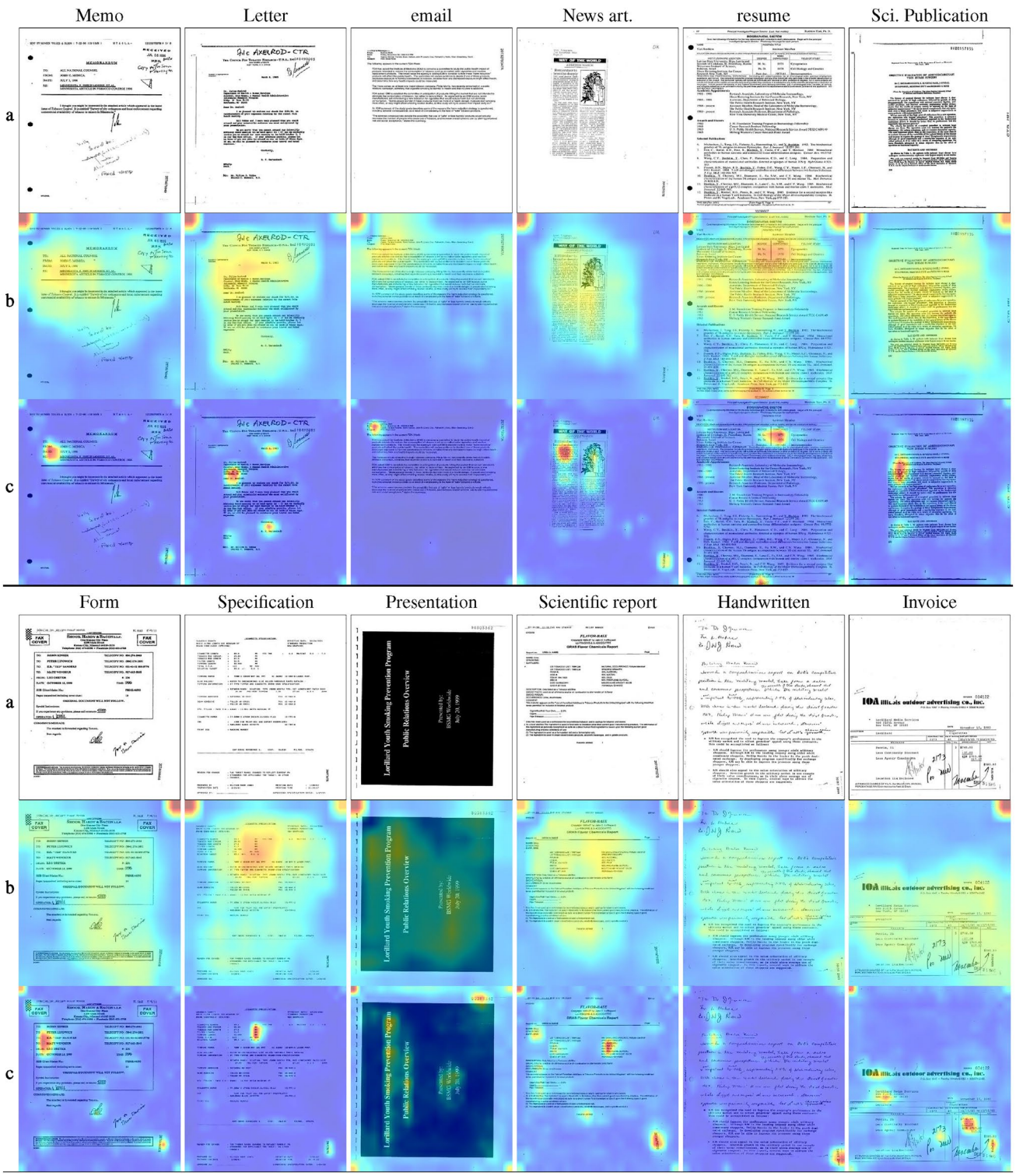

Fig. 5 Visually shows the importance of global feature vs. discriminative features for document image classification tasks. The 'a' rows show the sample images of RVL-CDIP dataset with category labels—Memo, Letter, Email, News article, Resume, Scientific Publication, Form, Specification, Presentation, Scientific report, Handwritten and Invoice. The 'b' rows illustrate the heat map calculated as the L_2 norm of the last convolutional layer before global average pooling, which is overlaid on the input image (i.e., global feature). The 'c' rows show the heat map calculated as the L_2 norm of the last convolutional layer before the discriminative feature head, which is overlaid on the input image (i.e., discriminative feature)

Table 2 Presents the document image classification accuracy on RVL-CDIP dataset, obtained by various methods based on the combination of images, text, and spatial features

Method	Accuracy(%)
Methods based on only visual feature	
Harley et al. [12]	89.80
SelfDoc [54]	90.49
Csurka et al. [14]	90.70
Tensmeyer et al. [13]	90.94
Afzal et al. [2]	90.97
Das et al. [1]	92.21
Sarkhel et al. [18]	92.77
Ours	**92.94**
Methods based on textual feature	
BERT-base [20]	89.81
UniLMv2-base [55]	90.06
LayoutLMv1-base [19]	91.78
Methods based on textual + spatial features	
UniLMv2-large [55]	90.20
LayoutLMv1-large [19]	91.90
Methods based on visual + textual + spatial features	
Single Modal [21]	93.03
Ensemble [21]	93.07
SelfDoc [54]	93.81
LayoutLMv1-large [19]	94.43
LayoutLMv2-large [56]	95.65
DocFormer [22]	**96.17**

Bold indicates best result

From Fig. 5, we observe that the global feature head gives importance to the top corners of the image irrespective of the category labels. While the discriminative feature head focuses on the meaning of full discriminating regions, which is essential to discriminate one particular category of the document from other classes. For example, the discriminative feature head concentrates on the "address part" for both Memo and Letter categories of documents while it focuses on "signature" only for Letter. In this case, "signature" is the discriminative feature to discriminate between Memo and Letter. The tabular structures generally occur in Form, Specification, Scientific report, and Invoice categories of documents and the discriminative feature head focuses on this tabular region (shown in Fig. 5). While the discriminative feature head fails to localize the handwriting region in the Handwritten category of documents. The encoding head overcomes this disadvantage by learning the textural pattern of the Handwritten type of documents.

State-of-the-Art Comparison

Table 2 presents a detailed comparison between the proposed method and the recent methods for document image classification tasks on RVL-CDIP [12] dataset. The works [1, 2, 12–14, 18] used only visual features for classifying document images. Among all existing methods using only visual features, Sarkhel et al. [18] obtained the highest accuracy (92.77%). While our method judicially combines visual features like global, textural, and local discriminative features for document classification and obtains an accuracy of 92.94% on RVL-CDIP dataset. The proposed method obtains state-of-the-art performance while using only the visual feature.

Methods like BERT-base [20], UniLMv2-base [55], and LayoutLMv1-base [19] uses the only textual feature for document classification. Among all these methods, LayoutLMv1-base [19] technique obtains the highest accuracy (91.78%). The proposed method using only visual features obtains 1.16% better accuracy than LayoutLMv1-base [19] method using only textual feature. Methods like UniLMv2-large [55] and LayoutLMv1-large [19] use both the textual and spatial features for document classification. Among them, LayoutLMv1-large [19] method obtains the highest accuracy (91.90%). Methods such as Single Modal [21], Ensemble [21], SelfDoc [54], LayoutLMv1-large [19], LayoutLMv2-large [56], and DocFormer [22] use visual, textual, and spatial features for document classification. Among all these methods, DocFormer [22] obtains state-of-the-art performance (96.17% accuracy).

Table 3 Shows the books' genre classification from their cover images in various settings, where $\oplus$ represents the addition the losses separately and $\otimes$ represents the concatenation of all in the last layers followed by a linear neural network

Feature	Mode	Accuracy↑
Global		35.10%
Global + discriminative	$\otimes$	35.62%
Global + encoding	$\otimes$	35.24%
Global + encoding + discriminative	$\otimes$	35.82%
Global + discriminative	$\oplus$	35.70%
Global + encoding	$\oplus$	35.32%
Global + encoding + discriminative	$\oplus$	**36.17**%

Bold indicates best result

Book Cover Classification

It is the task of identifying the genre of the book from its cover image. It is one of the challenging tasks in document image analysis [8] because books come with a wide variety of covers and styles, including nondescript and misleading covers. Unlike other object detection and classification tasks, genres are not concretely defined. Another problem is a large number of books that makes it unsuitable for exhaustive search methods.

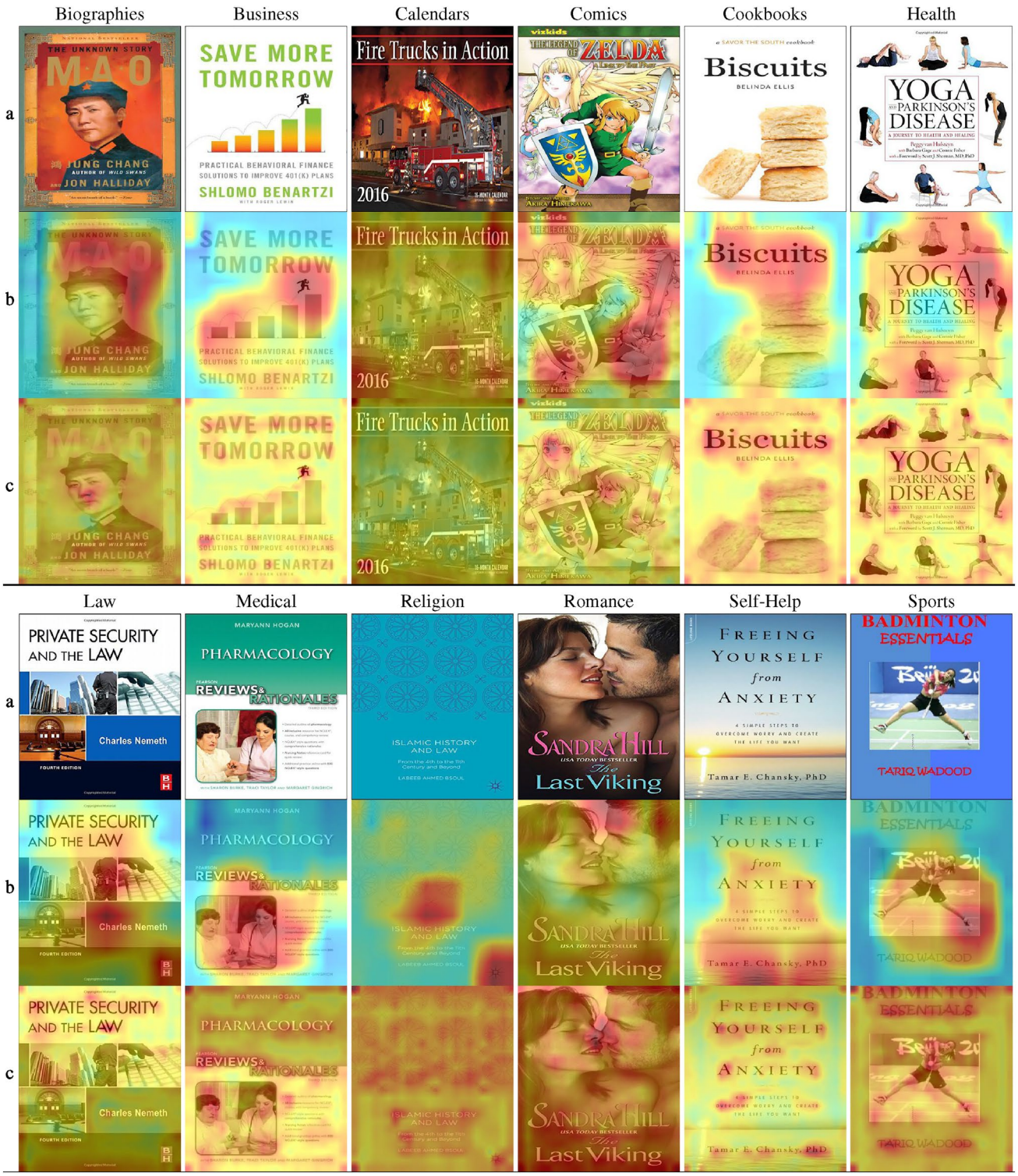

Fig. 6 Shows the importance of global and discriminative features for genre classification task. The 'a' rows show the sample images of Book-Cover dataset with category labels—Biographies, Business, Calendars, Comics, Cookbook, Health, Law, Medical, Religion, Romance, Self-Help and Sports. The 'b' rows illustrate the heat map calculated as the L_2 norm of the last convolutional layer before global average pooling, which is overlaid on the input image (i.e., global feature). The 'c' rows show the heat map calculated as the L_2 norm of the last convolutional layer before the discriminative feature head, which is overlaid on the input image (i.e., discriminative feature)

Table 4 Shows the class wise genre classification accuracy of Book-Cover dataset, obtained by various approaches

Genre	Classification accuracy↑			
	LeNet [8]	AlexNet [8]	FC-CNN+ FV-CNN [3]	Our
Arts and photography	05.80	12.10	17.37	**27.37**
Biographies and memoirs	05.30	13.20	17.37	**19.47**
Business and money	10.00	12.60	16.84	**28.95**
Calendars	18.90	47.90	44.74	**71.58**
Children's books	24.70	42.10	40.53	**44.74**
Comics and graphic novels	15.80	47.40	59.47	**67.89**
Computers and technology	29.50	44.70	51.58	**55.79**
Cookbooks food and wine	14.20	43.70	47.37	**56.84**
Crafts hobbies and home	07.40	17.40	30.00	**40.53**
Christian books and bibles	08.40	07.40	13.16	**18.42**
Engineering and transport	10.00	20.00	35.26	**36.84**
Health fitness and dieting	04.20	12.60	13.16	**18.95**
History	06.30	12.60	**25.79**	19.47
Humor and entertainment	05.30	10.50	11.58	**18.95**
Law	14.70	25.30	35.79	**40.53**
Literature and fiction	03.20	11.10	12.11	**17.37**
Medical books	12.60	19.50	25.79	**34.74**
Mystery thriller	23.70	34.20	36.84	**48.42**
Parenting and relationships	14.70	24.20	30.53	**31.58**
Politics and social sciences	03.70	06.80	11.58	**13.68**
Reference	13.20	20.00	23.68	**28.42**
Religion and spirituality	08.40	16.30	18.95	**29.47**
Romance	27.40	45.30	48.42	**56.32**
Science and math	08.40	14.20	**24.21**	23.68
Science fiction and fantasy	14.70	35.80	14.74	**41.05**
Self-help	13.70	14.20	**19.47**	15.79
Sports and outdoors	05.30	14.70	**32.11**	**32.11**
Teen and young adult	07.90	12.10	15.79	**27.37**
Test preparation	47.90	68.90	58.42	**73.68**
Travel	19.50	33.20	36.84	**45.16**
Total average	13.50	24.70	29.00	**36.17**

Bold indicates best result

Dataset and Pre-processing

We use the Book-Cover dataset [8], which contains 57K book cover images of 30 different genres. Each genre contains 1.9K images. Table 4 lists down all the genre categories. Figure 1a shows the sample images. This particular task is solved by the various global features extracted from deep neural networks such as LeNet [57] and AlexNet [23], and the results are reported in [8]. Even though our network takes input images of various sizes, we resize the input image to 227×227 to compare the result with the reported results of the existing approaches.

Ablation Study

Table 3 shows the ablation study on genre classification accuracy of various strategies. From the table, we observe that we obtain genre book classification accuracy of 35.10% only using the global feature. The use of the discriminative feature head and encoding feature head further improves the accuracy. The use of discriminative and encoding feature heads along with global feature head enhance the accuracy by 0.52% and 0.14%, respectively, in the case of concatenation strategy. The combination of all the three feature heads further improves the result by 0.72%. As we expect, the addition of feature heads loss gives the best result (36.17%) which is 0.35% better than the concatenating feature head strategy. We also observe that the classification accuracy using discriminative and global feature heads is better than encoding and global feature heads.

Learned Feature Visualization

We visualize the L_2 norm of the learned global and discriminative features, shown in Fig. 6. The global feature highlights the background region containing the word "MAO", but the discriminative feature highlights the face region for the cover pages of the category Biographies. A human face is ubiquitous to appear on the cover pages of books in the Biographies category. We observe that the discriminate feature head highlights the representative words of each category of book, which appear on the cover pages. A few examples are the discriminative feature head focuses on the words "SAVE MORE" in Business, year "2016" in Calendar, "BISCUITS" in Cookbooks, "LAW" in LAW, and "PHARMACOLOGY" in Medical categories of books.

State-of-the-Art Comparison

Table 4 presents a detailed comparison between the proposed and state-of-the-art approaches to book genre classification tasks. The proposed approach improves average accuracy by

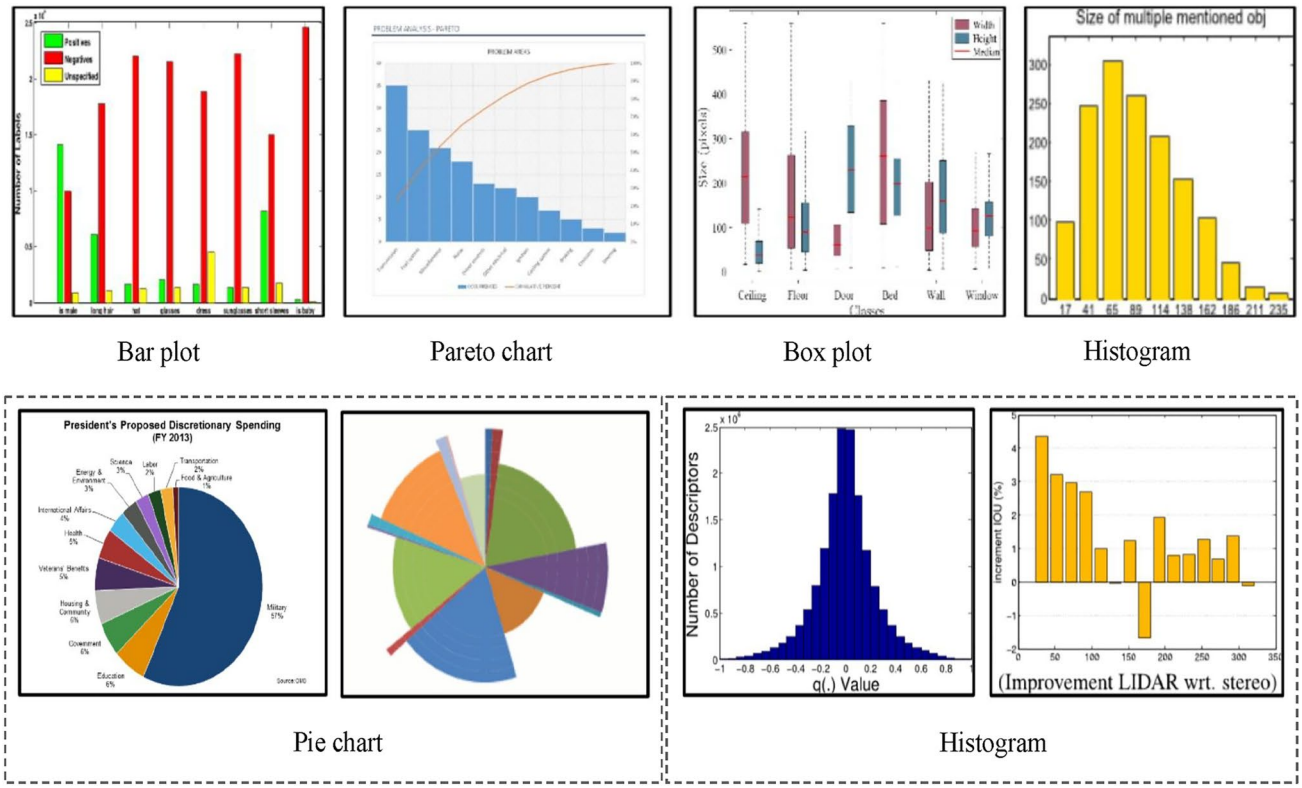

Fig. 7 Shows challenges on the DOCFIGURE dataset. The first row indicates images from different categories share similar visual properties. The second row highlights that the sample images from a specific type, like "Pie chart" and "Histogram" have a distinct structural property

22.67%, 11.47% and 7.17% over state-of-the-art methods—LeNet [8], AlexNet [8] and FC-CNN+FV-CNN [3], respectively. Compared to the proposed approach, FC-CNN+FV-CNN [3] feature outperforms in History, Science, Math, Self-help, & Sports and Outdoors classes. Iwana et al. [8] studied that the History category images have a high visual similarity with the images of other classes, such as Biographies & Memoirs, Politics & Social Sciences. The authors also pointed out that the high number of miss-classifications occurs in Biographies & Memoirs category.

Table 5 Presents the document figure classification accuracy in various settings, where ⊕ represents the addition of the losses separately, and ⊗ represents the concatenation of all in the last layers followed by a linear neural network

Feature	Mode	Accuracy↑
Global		95.91%
Global + discriminative	⊗	96.13%
Global + encoding	⊗	96.05%
Global + encoding + Discriminative	⊗	96.22%
Global + discriminative	⊕	96.15%
Global + encoding	⊕	96.08%
Global + encoding + discriminative	⊕	**96.24%**

Bold indicates best result

Document Figure Classification

It is a task of assigning category labels like Block diagram, Natural image, and Bar chart to the given document figure images. Classification of figures present in document images is complex due to inter-class visual similarity and intra-class visual dissimilarity. The existing methods [29, 33] using handcrafted features, fail to achieve good accuracy due to the extensive visual similarity among subcategories. Recently, a few techniques [36, 58] have been developed by convolutional neural networks to solve this problem.

Dataset

We use DOCFIGURE dataset [3] for this particular experiment. The dataset contains 32*K* images of 28 different categories of figures which are collected from scientific document images which correspond to scientific articles published in the CVPR, ECCV, and ICCV. conferences in last several years. The sample images are shown in Fig. 1d. Jobin et al. reported results of three baselines—FC-CNN, FV-CNN, and FC-CNN+FV-CNN in [3]. The DOCFIGURE dataset is the biggest dataset compare to the other document figure dataset Figureseer [36], Revision [33], Deepchart [58], and Karthikeyani and

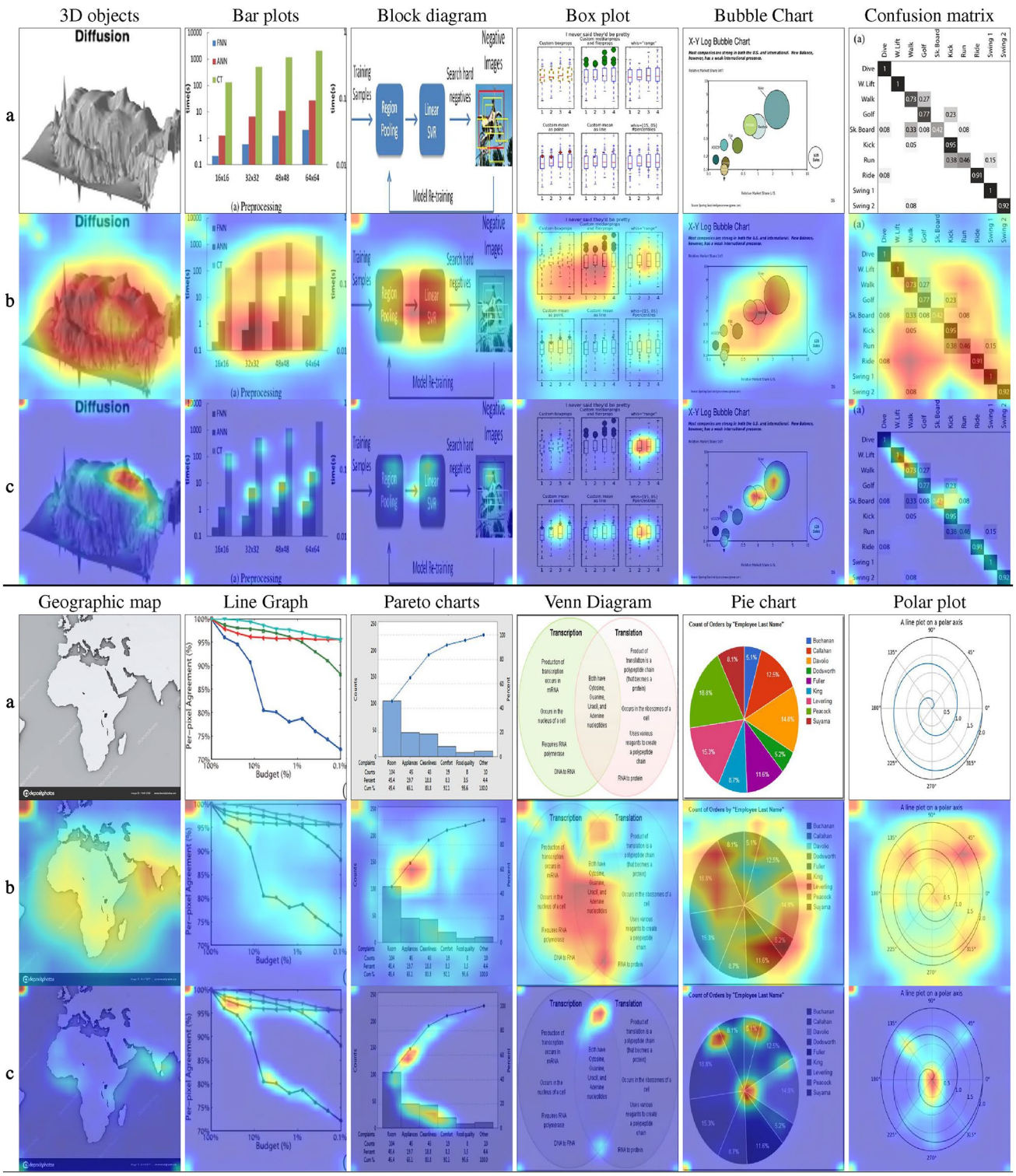

Fig. 8 Shows the importance of global features vs. discriminative features for document figure classification tasks. The 'a' rows show the sample images of DocFigure dataset [3] with selected category labels—3D object, Bar chart, Block diagram, Box plot, Bubble chart, Confusion matrix, Geographic map, Line graph, Pareto chart, Venn diagram, Pie chart, and Polar chart. The 'b' rows illustrate the heat map calculated as the L_2 norm of the last convolutional layer before global average pooling, which is overlaid on the input image (i.e., global feature). The 'c' rows show the heat map calculated as the L_2 norm of the last convolutional layer before the discriminative feature head, which is overlaid on the input image (i.e., discriminative feature)

Table 6 Shows the class wise document figure classification accuracy of DocFigure dataset, obtained by various approaches

Labels	Classification Accuracy↑			
	FC-CNN [3]	FV-CNN [3]	FV-CNN+ FC-CNN	Our
3D object	98.24%	94.73%	**98.53%**	97.81%
Algorithm	93.81%	91.75%	93.81%	**96.77%**
Bar chart	**93.97%**	91.97%	93.64%	93.10%
Box plot	91.39%	88.07%	**92.95%**	92.05%
Flow chart	92.53%	91.04%	91.38%	**97.01%**
Heat map	99.25%	95.89%	96.27%	**99.62%**
Histogram	**94.89%**	88.26%	**94.89%**	91.69%
Medical image	97.87%	92.55%	**98.93%**	96.90%
Pie chart	91.66%	89.81%	94.44%	**97.69%**
Polar chart	85.71%	78.57%	85.71%	**89.55%**
Area chart	84.61%	91.02%	92.30%	**100.00%**
Block diagram	97.26%	97.65%	**98.43%**	91.93%
Bubble chart	80.95%	91.66%	90.47%	**97.78%**
Confusion matrix	85.22%	89.65%	**93.10%**	91.36%
Contour plot	59.34%	74.72%	72.52%	**91.67%**
Geographic map	88.59%	95.81%	95.43%	**98.57%**
Line graph	98.49%	98.84%	**99.33%**	98.50%
Mask	**99.23%**	**99.23%**	**99.23%**	98.34%
Natural image	98.04%	98.25%	**99.23%**	98.57%
Pareto chart	87.17%	96.15%	**97.43%**	95.97%
Radar chart	78.94%	86.84%	85.52%	**90.00%**
Scatter plot	90.14%	91.19%	**93.66%**	89.89%
Sketches	95.65%	96.37%	**98.18%**	94.78%
Surface plot	76.76%	89.89%	88.88%	**93.59%**
Tables	97.25%	98.73%	97.67%	**99.34%**
Tree diagram	67.04%	68.18%	70.45%	**82.14%**
Vector plot	79.86%	81.94%	86.80%	**94.32%**
Venn diagram	87.03%	93.51%	93.05%	96.00%
Average accuracy	88.96%	90.80%	92.90%	**96.24%**

Bold indicates best result

Nagarajan [59]. The Fig. 7 illustrates the inter-similar and intra-dissimilar visual properties of the DocFigure dataset.

Ablation Study

Table 5 shows the ablation study for solving document figure classification tasks using various settings. Using the global feature alone gives an accuracy of 95.91% on the document figure classification task. It also shows that combining the discriminative feature with the global feature yields better results than combining the encoding feature with the global feature in concatenation and addition settings. Combining all three—discriminative, encoding, and global features gives the best classification accuracy (96.24% and 96.22%, respectively) for both scenarios.

Learned Feature Visualization

We visualize L_2 norms of global and discriminative features (shown in Fig. 8) to get an insight and analyze the importance of the feature for document figure classification tasks. From Fig. 8, we observe that the global feature head focuses on complete object region while the discriminative feature head highlights only essential parts of the object region. These essential parts of the object region are vital for recognizing one category of figure from other categories. We noticed that discriminative feature head focuses on (1) a 3D shaped region in 3D Object, (2) the trips of the bar in Bar chart, (3) bubble circle in Bubble chart, (4) lines in Line graph, and (5)"India" region in Geographic map categories of document figures. Figure 8 highlights the discriminative head learns features corresponding to the meaningful region of the input figure image.

State-of-the-Art Comparison

Table 6 shows the class wise accuracy of DocFigure dataset using various methods. The proposed method obtains improved accuracy over the state-of-the-art techniques—FC-CNN, FV-CNN, and FC-CNN+FV-CNN for 15 categories of document figures. The proposed method obtains the maximum accuracy improvement (i.e., 19.15%) of the Contour plot over state-of-the-art techniques. The proposed technique also improve average classification accuracy by 7.28%, 5.44%, and 3.34% over FC-CNN, FV-CNN, and FV-CNN+FC-CNN, respectively.

Script Identification in Multi-lingual Document Images

It is the task of identifying scripts in multi-lingual document images. It has a wide range of applications including automatic storage of multi-script document images, document image retrieval, video indexing and retrieval, and document sorting in digital libraries [4]. Features play an important role in the script identification system. The article [4] summarizes the various feature categories popularly applied in the script identification techniques. Various texture features like Gabor filter, gray level co-occurrence matrix and wavelet are considered for script identification in multi-script documents [60–68]. Due to generalization capability, recently, deep features are considered to identify script in multi-lingual document images [69].

Fig. 9 Shows sample images of CVSI-2015 dataset. Each row of the image represent the sample images of **a** Arabic, **b** Bengali, **c** English (Roman), **d** Gujarathi, **e** Hindi (Devnagari), **f** Kannada, **g** Oriya, **h** Punjabi (Gurumukhi), **i** Tamil, and **j** Telegu, respectively

Dataset

We use CVSI-2015 dataset [70] for this particular experiment. The dataset is composed of images from news videos in various Indian languages. It contains 6412 training text images and 3207 test text images from 10 different scripts, namely Arabic, Bengali, English, Gujarati, Hindi, Kannada, Oriya, Punjabi, Tamil, and Telugu. The sample images of ten languages are shown in Fig. 9. To handle the arbitrary size of the word image and enhance the unique curves in the script, we perform the following pre-processing steps: (i) conversion of all color images to grayscale images in which the character's areas are darker than the background, (ii) re-scale the image with widths 100, 40, 80 and 160, respectively, by keeping the aspect ratio constant, and (iii) arrange this scaled images in a 384×384 canvas as shown in the second row and sixth row in the Fig. 10.

Ablation Study

Table 7 shows the ablation study on the script identification task under various network configurations. The global feature obtains 97.00% accuracy. In this particular problem, the encoded i.e., texture feature is more effective than the discriminative feature, as indicated in Table 7. The encoded feature combined with the global feature obtains 0.5% and 0.57% improved accuracy over the combination of discriminative and global features under concatenation and addition settings. The classification accuracy is further improved using a combination of global, discriminative and encoded features in both strategies.

Learned Feature Visualization

We visualize the heat map calculated as the L_2 norm of the last convolutional layer of global average pooling and the convolutional layer before the discriminative feature head, shown in Fig. 10. From the figure, we notice that the global feature head focuses on words of various scales. However, the discriminative feature head fails to learn a discriminative patch from the scrips, e.g., Arabic, Gujarati, Oriya, and Telugu.

State-of-the-Art Comparison

We compare the performance of our method with state-of-the-art techniques on the script classification task. These results are summarized in Table 8. The table highlights that the proposed approach obtains a very close result to Google.

Ablation Study on Hyperparameters

The more effective hyperparameters in the proposed architecture are the number of codewords (K) in the encoding feature head and the number of channels per class in the CCP layer (m) of the discriminative feature head. We conduct a


Arabic
Bengali
English
Gujarati
Hindi
a
b
c
d
Kannada
Oriya
Punjabi
Tamil
Telugu
a
b
c
d

◂**Fig. 10** Shows the importance of global vs. discriminative features for document figure classification tasks. The 'a' rows show the sample images of CVSI dataset [70] with languages—Arabic, Bengali, English, Gujarati, Hindi, Kannada, Oriya, Punjabi, Tamil, and Telugu. The 'b' rows show the pre-processed word image before being fed into the classification network. The 'c' rows illustrate the heat map calculated as the L_2 norm of the last convolutional layer before global average pooling, which is overlaid on the input image (i.e., global feature). The 'd' rows show the heat map calculated as the L_2 norm of the last convolutional layer before the discriminative feature head, which is overlaid on the input image (i.e., discriminative feature)

study to determine the optimum value of K and m for four classification problems.

In the study of codewords (K), we use the architecture having global and encoding feature heads with the addition of the losses. We calculate the classification accuracy of four tasks—document image classification, book cover classification, document figure classification and script classification corresponding to four datasets—RVL-CDIP, Book cover, DOCFigure, and CVSI with varying K from 4 to 256. Figure 11 shows the variation on accuracy with the change of K. From the figure, we observe that the variation in accuracy is very less with the change in K for all tasks (almost flat curve). We also observe from the figure that the best value of K for each dataset directly relates to the total number of classes present in the dataset. The CVSI dataset has 10 classes and the best K is 8 for this dataset. The RVL-CDIP dataset has 16 classes and the best K is 16. The Book cover and DOCFigure datasets have 30 and 28 classes, respectively, the best K is 32 for both datasets.

In the study of the number of channels per class in the CCP layer (m), we use the architecture with global and discriminative feature heads with the addition of the losses. First, we calculate the classification accuracy of four classification tasks with m value ranging from 5 to 35. Figure 12 shows the variation on accuracy with the changing m for four different tasks corresponding to four datasets. From the figure, we observe that the best accuracy is obtained with the value of K is equal to 20, irrespective of the dataset.

Table 7 Shows the script identification accuracy in various settings, where ⊕ represents the addition of the losses separately, and ⊗ represents the concatenation of all in the last layers followed by a linear neural network

Features	Mode	Accuracy↑
Global		97.00%
Global + discriminative	⊗	97.71%
Global + encoding	⊗	98.21%
Global + encoding + discriminative	⊗	98.64%
Global + discriminative	⊕	97.78%
Global + encoding	⊕	98.35%
Global + encoding + discriminative	⊕	**98.83%**

Bold indicates best result

Table 8 Shows the script identification accuracy of CVSI-2015 dataset [70], obtained by various approaches

Approach	Accuracy↑
C-DAC	84.66%
CUK	74.06%
HUST	96.69%
CVC-1	95.88%
CVC-2	96.00%
Shi et al. [10]	96.70%
Sing et al. [71]	98.13%
Our	98.83%
Google	**98.91%**

Bold indicates best result

Conclusion

We introduce a deep multi-modular feature extraction architecture for various classification tasks in document image analysis. The proposed architecture extracts three features—discriminative, encoded/texture, and global. Diverse experiments conclude that a combination of discriminative, texture, and global features performs better for various classification tasks—document image classification, genre classification of book, document figure classification, and script identification. We visualize the L_2 norm of the learned global and discriminative features in the form of the heat map, highlighting their importance for various classification tasks. The heat maps highlight that discriminative features are more useful for document image classification, document figure classification, and genre classification than other

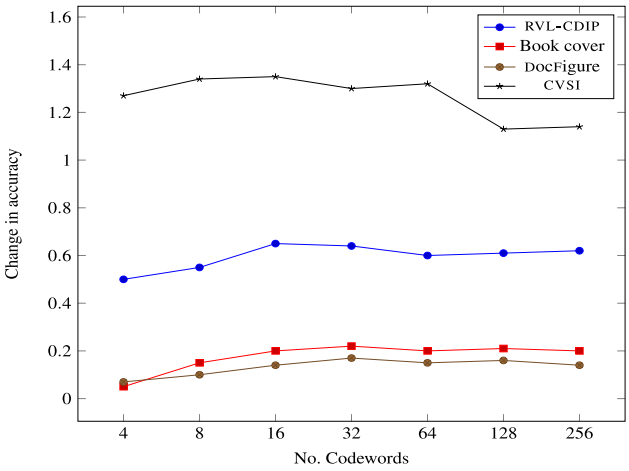

Fig. 11 Shows the variation on accuracy with the changes in number of codewords K for four tasks—document image classification, book cover classification, document figure classification, and script classification corresponding to four datasets—RVL-CDIP, Book cover, DOCFigure, and CVSI. Here, we combine the global feature head loss with the encoding feature head loss

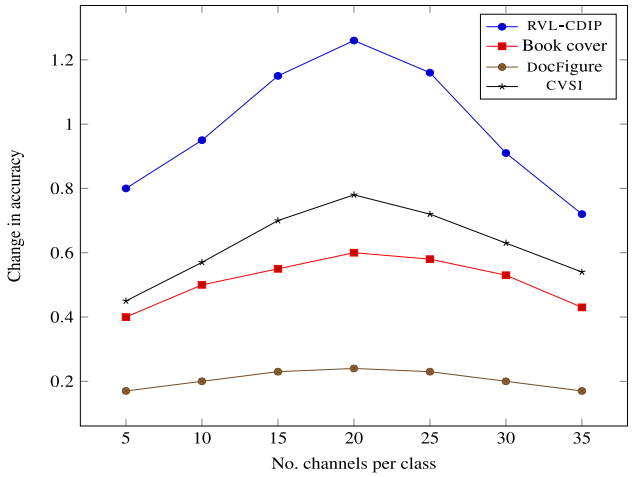

Fig. 12 Shows the variation on accuracy with the changes in number of channels per class in the CCP layer (m) for four tasks—document image classification, book cover classification, document figure classification and script classification corresponding to four datasets—RVL-CDIP, Book cover, DOCFIGURE, and CVSI. Here, we combine the global feature head loss with the discriminating feature head loss

features. In contrast, the encoded feature is more essential than other features for the script identification task. In the future, we will explore the proposed architecture for classroom slide retrieval and signature verification.

Funding One of the authors, Jobin K.V., received a Visvesvaraya Ph.D. fellowship from the government of India.

Declarations

Conflict of interest All authors declare that they have no conflicts of interest.

Ethics approval This article does no contain any studies with human participants or animals performed by any of the authors.

References

1. Das A, Roy S, Bhattacharya U, Parui SK. Document image classification with intra-domain transfer learning and stacked generalization of deep convolutional neural networks. In: ICPR 2018.
2. Afzal MZ, Kolsch A, Ahmed S, Liwicki M. Cutting the error by half investigation of very deep cnn and advanced training strategies for document image classification. In: ICDAR 2017.
3. Jobin K, Mondal A, Jawahar C. Docfigure: a dataset for scientific document figure classification. In: GREC 2019.
4. Ubul K, Tursun G, Aysa A, Impedovo D, Pirlo G, Yibulayin T. Script identification of multi-script documents: a survey. IEEE Access. 2017.
5. Torkkola K. Discriminative features for text document classification. Formal Pattern Anal Appl. 2004;6(4):301–8.
6. Jiang H, Pan Z, Hu P. Discriminative learning of generative models: large margin multinomial mixture models for document classification. Pattern Anal Appl. 2015;18(3):535–51
7. Soleimani H, Miller DJ. Exploiting the value of class labels on high-dimensional feature spaces: topic models for semi-supervised document classification. Pattern Anal Appl. 2019;22(2):299–309
8. Iwana BK, Rizvi STR, Ahmed S, Dengel A, Uchida S. Judging a book by its cover. 2016.
9. Singh AK, Mishra A, Dabral P, Jawahar CV. A simple and effective solution for script identification in the wild. Pattern Recognit. 2016. p. 428–33
10. Shi B, Bai X, Yao C. Script identification in the wild via discriminative convolutional neural network. Pattern Recognit. 2016;52:448–58
11. Liu L, Wang Z, Qiu T, Chen Q, Lu Y, Suen CY. Document image classification: progress over two decades. Neurocomputing. 2021;453:223–40.
12. Harley AW, Ufkes A, Derpanis KG. Evaluation of deep convolutional nets for document image classification and retrieval. In: ICDAR 2015.
13. Tensmeyer C, Martinez T. Analysis of convolutional neural networks for document image classification. In: ICDAR 2017.
14. Csurka G, Larlus D, Gordo A, Almazan J. What is the right way to represent document images? 2016.
15. Wang Y, Morariu VI, Davis LS. Learning a discriminative filter bank within a cnn for fine-grained recognition. In: CVPR 2018.
16. Zheng H, Fu J, Zha Z-J, Luo J. Looking for the devil in the details: learning trilinear attention sampling network for fine-grained image recognition. In: CVPR 2019.
17. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-first AAAI conference on artificial intelligence 2017.
18. Sarkhel R, Nandi A. Deterministic routing between layout abstractions for multi-scale classification of visually rich documents. In: IJCAI 2019.
19. Xu Y, Li M, Cui L, Huang S, Wei F, Zhou M. Layoutlm: pre-training of text and layout for document image understanding. In: ACM SIGKDD international conference on knowledge discovery & data mining 2020.
20. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018.
21. Dauphinee T, Patel N, Rashidi M. Modular multimodal architecture for document classification. 2019.
22. Appalaraju S, Jasani B, Kota BU, Xie Y, Manmatha R. Docformer: end-to-end transformer for document understanding. In: ICCV 2021.
23. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: NIPS; 2012. p. 1097–105.
24. Zujovic J, Gandy L, Friedman S, Pardo B, Pappas TN. Classifying paintings by artistic genre: an analysis of features & classifiers. In: International workshop on multimedia signal processing 2009.
25. Chiang H, Ge Y, Wu C. Classification of book genres by cover and title. Computer science: class report; 2015.
26. Biradar GR, Raagini J, Varier A, Sudhir M. Classification of book genres using book cover and title. In: International conference on intelligent systems and green technology (ICISGT) 2019.
27. Lucieri A, Sabir H, Siddiqui SA, Rizvi STR, Iwana BK, Uchida S, Dengel A, Ahmed S. Benchmarking deep learning models for classification of book covers. SN computer science 2020.
28. Liu Y, Lu X, Qin Y, Tang Z, Xu J. Review of chart recognition in document images. In: VDA 2013.
29. Zhou YP, Tan CL. Hough technique for bar charts detection and recognition in document images. In: ICIP 2000.

30. Zhou YP, Tan CL. Bar charts recognition using hough based syntactic segmentation. In: ICTAD 2000.
31. Zhou Y, Tan CL. Learning-based scientific chart recognition. In: IWGR 2001.
32. Prasad VSN, Siddiquie B, Golbeck J, Davis LS. Classifying computer generated charts. In: CBMI 2007.
33. Savva M, Kong N, Chhajta A, Fei-Fei L, Agrawala M, Heer J. Revision: automated classification, analysis and redesign of chart images. In: User interface software and technology 2011.
34. Kavasidis I, Palazzo S, Spampinato C, Pino C, Giordano D, Giuffrida D, Messina P. A saliency-based convolutional neural network for table and chart detection in digitized documents. 2018.
35. Tang B, Liu X, Lei J, Song M, Tao D, Sun S, Dong F. Deepchart: combining deep convolutional networks and deep belief networks in chart classification. Signal Process. 2016.
36. Siegel N, Horvitz Z, Levin R, Divvala S, Farhadi A. Figureseer: parsing result-figures in research papers. In: ECCV 2016.
37. Aletras N, Mittal A. Labeling topics with images using a neural network. In: European conference on information retrieval 2017.
38. Charbonnier J, Sohmen L, Rothman J, Rohden B, Wartena C. Noa: a search engine for reusable scientific images beyond the life sciences. In: European conference on information retrieval 2018.
39. Shijian L, Tan CL. Script and language identification in noisy and degraded document images. In: IEEE Transactions on PAMI 2007.
40. Zhou L, Lu Y, Tan CL. Bangla/English script identification based on analysis of connected component profiles. In: International workshop on document analysis systems 2006.
41. Sharma N, Pal U, Blumenstein M. A study on word-level multi-script identification from video frames. In: 2014 international joint conference on neural networks (IJCNN) 2014.
42. Mei J, Dai L, Shi B, Bai X. Scene text script identification with convolutional recurrent neural networks. In: ICPR 2016.
43. Lu L, Yi Y, Huang F, Wang K, Wang Q. Integrating local cnn and global cnn for script identification in natural scene images. IEEE Access. 2019;7:52669–79.
44. Bhunia AK, Konwer A, Bhunia AK, Bhowmick A, Roy PP, Pal U. Script identification in natural scene image and video frames using an attention based convolutional-LSTM network. Pattern Recognit. 2019;85:172–84.
45. Ghosh M, Mukherjee H, Obaidullah SM, Santosh K, Das N, Roy K. Lwsinet: a deep learning-based approach towards video script identification. Multim Tools Appl. 2021;80(19):29095–128.
46. Chatfield K, Simonyan K, Vedaldi A, Zisserman A. Return of the devil in the details: delving deep into convolutional nets 2014. arXiv:1405.3531
47. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014.
48. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: CVPR 2016.
49. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: CVPR 2016.
50. Cimpoi M, Maji S, Vedaldi A. Deep filter banks for texture recognition and segmentation. In: CVPR 2015.
51. Zhang H, Xue J, Dana K. Deep ten: texture encoding network. In: CVPR 2017.
52. Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid scene parsing network. In: CVPR 2017.
53. Loshchilov I, Hutter F. Sgdr: stochastic gradient descent with warm restarts. 2016.
54. Li P, Gu J, Kuen J, Morariu VI, Zhao H, Jain R, Manjunatha V, Liu H. Selfdoc: self-supervised document representation learning. In: CVPR 2021.
55. Bao H, Dong L, Wei F, Wang W, Yang N, Liu X, Wang Y, Gao J, Piao S, Zhou M, et al. Unilmv2: Pseudo-masked language models for unified language model pre-training. In: International conference on machine learning 2020. PMLR.
56. Xu Y, Xu Y, Lv T, Cui L, Wei F, Wang G, Lu Y, Florencio D, Zhang C, Che W, et al. Layoutlmv2: multi-modal pre-training for visually-rich document understanding. 2020.
57. LeCun Y, Bottou L, Bengio Y, Haffner P, et al. Gradient-based learning applied to document recognition. In: Proceedings of the IEEE. 1998.
58. Tang B, Liu X, Lei J, Song M, Tao D, Sun S, Dong F. Deepchart: combining deep convolutional networks and deep belief networks in chart classification. Signal Process. 2016;124:156–61.
59. Karthikeyani V, Nagarajan S. Machine learning classification algorithms to recognize chart types in portable document format (pdf) files. Int J Comput Appl. 2012;39(2):1–5.
60. Busch A, Boles WW, Sridharan S. Texture for script identification. IEEE Transactions on PAMI 2005.
61. Busch A. Multi-font script identification using texture-based features. In: Campilho A, Kamel M (eds) Image analysis and recognition. 2006.
62. Singhal V, Navin N, Ghosh D. Script-based classification of handwritten text documents in a multilingual environment. In: RIDE-MLIM 2003.
63. Jaeger S, Ma H, Doermann D. Identifying script on word-level with informational confidence. In: ICDAR 2005.
64. Pati PB, Ramakrishnan A. Word level multi-script identification. Pattern Recognit Lett. 2008;29(9):1218–29.
65. Kunte RS, Samuel RDS. On separation of Kannada and English words from a bilingual document employing Gabor features and radial basis function neural network. ICCR 2005.
66. Philip B, Samuel RS. A novel bilingual OCR for printed Malayalam-English text based on Gabor features and dominant singular values. In: ICDIP 2009.
67. Rani R, Dhir R, Lehal GS. Script identification of pre-segmented multi-font characters and digits. In: ICDAR 2013.
68. Chanda S, Franke K, Pal U. Identification of indic scripts on torn-documents. In: ICDAR 2011.
69. Ukil S, Ghosh S, Md Obaidullah S, Santosh KC, Roy K, Das N. Deep learning for word-level handwritten indic script identification. CoRR 2018.
70. Sharma N, Mandal R, Sharma R, Pal U, Blumenstein M. ICDAR2015 competition on video script identification (CVSI 2015). In: ICDAR 2015.
71. Singh AK, Mishra A, Dabral P, Jawahar C. A simple and effective solution for script identification in the wild. In: DASW 2016.