Two Stream Deep Network for Document Image Classification

Muhammad Nabeel Asim*†, Muhammad Usman Ghani Khan†, Muhammad Imran Malik‡,
Khizar Razzaque†, Andreas Dengel*, Sheraz Ahmed*
Email: firstname.lastname@dfki.de
*German Research Center for Artificial Intelligence (DFKI), 67663 Kaiserslautern, Germany
†National Center for Artificial Intelligence (NCAI), University of Engineering and Technology, Lahore, Pakistan
‡National Center for Artificial Intelligence (NCAI), National University of Sciences and Technology, Islamabad, Pakistan

I. Abstract

This paper presents a novel two-stream approach for document image classification. The proposed approach leverages textual and visual modalities to classify document images into ten categories, including letter, memo, news article, etc. In order to alleviate dependency of textual stream on performance of underlying OCR (which is the case with general content based document image classifiers), we utilize a filter based feature-ranking algorithm. This algorithm ranks the features of each class based on their ability to discriminate document images and selects a set of top 'K' features that are retained for further processing. In parallel, the visual stream uses deep CNN models to extract structural features of document images. Finally, textual and visual streams are concatenated together using an average ensembling method. Experimental results reveal that the proposed approach outperforms the state-of-the-art system with a significant margin of 4.5% on publicly available Tobacco-3482 dataset.

***Index Terms*—Document Image Classification, Filter based feature selection, Multi-Channel CNN, RVL-CDIP, Tobacco-3482, Inception V3**

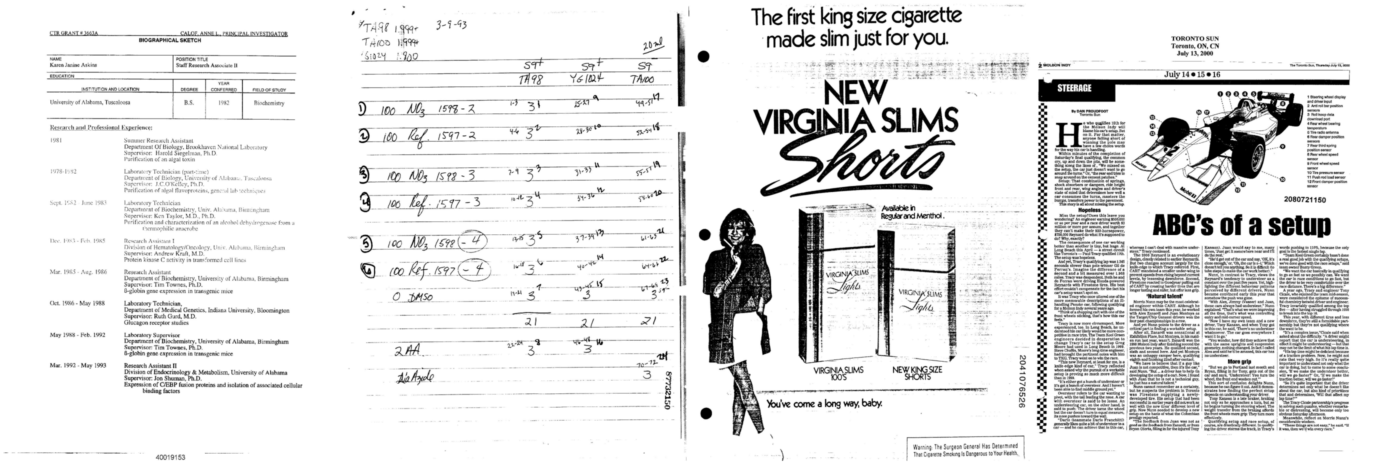
II. Introduction

Text document image classification plays an important part in multifarious information retrieval and text recognition tasks performed by diverse document analysis and processing systems. Text document image classification methodologies are categorized into structural and content based approaches. In computer vision, with the invention and huge success of AlexNet in 2012 [1], deep learning attracted researchers to develop deeper architectures for multifarious tasks of computer vision, natural language processing, and speech recognition. Lately, several Convolutional Neural Network (CNN) architectures have been proposed for image classification such as ZFNet[2], GoogleNet [3], VGNet [4], and ResNet [5]. Moreover, Convolutional Neural Networks have also witnessed a significant advancements in various aspects such as Convolutional layers, activation functions, pooling layers, loss functions, optimization, and regularization. Besides this, transfer learning also played an integral role to raise the performance of text document image classification. For instance, Afzal et al. [6] used a pretrained network on gigantic image dataset (ImageNet) for the task of text document image classification. Their experimental results showed that transfer learning significantly improved the performance of classification, although images of ImageNet [7] dataset (shown in Figure 1) were totally different from the images of Tobacco-3482 dataset (revealed in Figure 2).

The state-of-the-art methods for document classification utilize only visual information to classify text document images and face the problem of low inter-class and high intra-class structural variations of text document images [8] (Figure 3 shows visual structural similarities among different classes). Thus, these approaches fail to distinguish the sample images of highly corelated classes. Computationally to illustrate this problem more effectively, we used InceptionV3 model to extract fixed length vectors representing structural features of text document images present in Tobacco-3482 dataset. Afterwards, all feature vectors of each class are averaged into single feature vector, and finally similarity among averaged feature vectors is computed using cosine similarity. Amongst all classes of Tobacco-3482 dataset, three classes namely Letter, Memo, and Report got the highest similarity score. Figure 3 shows the sample images of highly co-related classes (Letter, Memo, Report). As illustrated by Figure 3, considering the document structure, each document of one class match the layout properties of one or more documents present in other classes, for instance, first document of Letter class is extremely similar to the first document of Memo and Report classes and second document of Letter class imitate the structure of second documents present in rest of the classes.

In order to tackle the problem of high inter-class similarity, Noce et al. [8] proposed a novel text document image classification methodology where they initially extracted text from document images using an OCR, followed by the ranking of textual features using Penas et al. [9] feature weighting formula. Ranked features were embedded again within the images in form of colors where each color represents a certain class. They reported that the proposed methodology raised the accuracy by 5.6% on the Tobacco-3482 dataset. Text document classification has witnessed a significant improvement with the invention of advance filter based feature selection approaches [10]. In this paper, we propose a two-stream classification

Fig. 1: Samples of ImageNet dataset



BIOGRAPHICAL SKETCH
The first king size cigarette made slim just for you.
NEW VIRGINIA SLIMS Shorts
You've come a long way, baby.
TORONTO SUN
ABC's of a setup

Fig. 2: Samples of Tobacco 3482 dataset

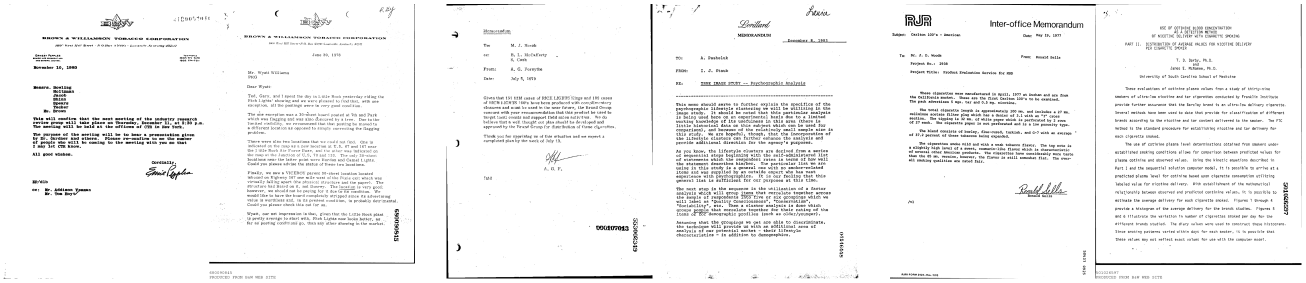
methodology for text document image classification. In particular, the contributions of this paper are:

1) Leveraging the potential of filter-based feature ranking algorithm
2) Employment of a multi-channel CNN for document text classification
3) A novel merging scheme for the fusion of both visual and textual cues using average ensembling

Our main contribution is the embedding of a better filter-based feature ranking algorithm in the state-of-the-art hybrid text document image classification methodologies. OCR does not extract the text accurately for classes like advertisement, and hand written, thus affects the performance of classification model. However, the embedded filter-based feature ranking metric reduces the network's reliance on poor-quality features, resulting in good performance even for cases where the OCR system fails. Moreover, different word embedding algorithms embed either syntactic or semantic features. In order to utilize both features effectively, a multi-channel CNN is used. Despite the advantages of pretrained word embeddings, the utility in our case is limited due to OCR errors which results in approximately a hit-rate of 25% due to errors in the transcription. In order to leverage these word embeddings, even in cases where the transcription is poor, we leverage a multi-channel CNN, where we stack the output of the Word2Vec model on the first channel, while the rest of the channels are randomly initialized. This allows the network to cater for the rest of the 75% of the cases, where the embedding could not be obtained due to these transcription errors. We leverage both textual and visual cues in our experimentation.

III. Related Work

Researchers have used a variety of features like bag of words (BoW), document structure, font sizes, column structures, and occurrence of text and non-text regions for classification of text document images [11], [12], [13], In the past, region based algorithms have shown intriguing results for structure dependant classes like letters and forms. Local image analysis is also adopted for text document image classification [14], [15]. Moreover, content of a text document image is also used for classification tasks [16]. This section provides a bird's eye view on the state-of-the-art approaches proposed for text document image classification. Kumar et al. [17] presented a methodology which depends on codewords extracted from different patches of text document images. Lately, Kumar et al. [18] came up with another approach which constructed a codebook having SURF descriptors of underlay text document images and used the developed codebook for classification.



(a) Letter (b) Memo (c) Report

Fig. 3

Chen et al. [19] used low level features of underlay images in order to classify structured documents. Reddy and Govindaraja [20] utilized pixel data of binary images in order to classify form documents.

Amongst all, convolutional neural networks (CNNs) based approaches have achieved a huge success in text document image classification. LeCun et al. [21] compared multifarious classification methodologies to extrapolate that CNNs handle the variability of 2D images extremely well and outshine all existing methodologies. For text document image classification, Kang et al. [22] were the first to utilize CNNs on Tobacco 3482 benchmark dataset. They used a shallow convolutional neural network to surpass the performance of trivial structure based approaches. Likewise, Afzal et al. [6], and Harley et al. [23] facilitated a major breakthrough as they revealed that transfer learning can be applied for text document image classification. Moreover, Harley et al. [23] developed a dataset namely RVL-CDIP having 400000 documents of 16 diverse classes. This proved a notable contribution as it allowed the evaluation of several neural network based approaches. Joutel et al. [24] used a very deep convolutional neural network in order to categorize 1.2 million 2D images of 1000 diverse classes. They extrapolated that CNN learned representation was transferable for different tasks. Furthermore, Girshick et al. [25] revealed that, in case of scarce data, classification performance gets improved when supervised pre-training on large data, and fine tuning on small data are performed. Kolsch et al. [26] proposed a methodology to resolve the problems of optimal feature extraction and long training time. They utilized CNN for feature extraction and extreme learning machines (ELMs) for classifying text document images into predefined classes.

Many researchers and practitioners have done notable work to improve the performance of text document classification, however, classifying text document images using the content of underlay images has not been explored significantly. The state-of-the-art text document classification work that can be adopted for text document image classification is discussed in this section. For example, multi-channel CNN proposed by Zhang et al. [27] utilized different word embeddings at different channels and increase the classification performance of the model by joining all channels to create a final feature vector. Zhang et al. [28] used a deep CNN for text classification that consists of six convolution and three fully connected layers. A very deep CNN was proposed and utilized by Conneau et al. [29] that have 29 convolution layers. Only small convolution and pooling layers were used in this deep architecture. This was the first time when someone used a neural network of such depth for text classification. A hybrid approach namely recurrent convolutional neural network was proposed by Lai et al. [30]. In this approach, a recurrent neural network (RNN) is followed by a CNN where RNN captures the contextual information and CNN extracts the features that play key role in text classification.

IV. Methodology

This sections illustrates proposed two stream methodology of text document image classification. In the first stream, text document images are fed to an InceptionV3 [31] model, whereas in the second stream, firstly text document images are converted into textual documents using an OCR then all the unique features of textual data are ranked using filter based feature ranking algorithm (ACC2) in order to feed a vocabulary of top k features to the embedding layer of multi-channel convolutional neural network (CNN). Finally, both streams are concatenated using average ensembling approach for the task of text document image classification. Figure 4 reveals the phases of proposed text document image classification methodology which are discussed in following subsections.

A. *Visual Stream*

In order to process images at first stream, all the images are downsized to 299x299 dimensions where every color channel is zero centered with respect to ImageNet dataset just to allow the model to converge faster. In our experimentation, we utilize InceptionV3 model to classify text document images using transfer learning. We have trained InceptionV3 [31] on RVL-CDIP dataset using ImageNet weights and utilized transfer learning to classify tobbacco-3482 text dcoument images. InceptionV3 [31] is a very deep CNN architecture consists of 42 layers. Its architecture consists of different inception modules stacked linearly. For effective down sampling of feature maps, two efficient grid size reduction modules are utilized in the InceptionV3 architecture. The feature maps of the final inception module are pooled using global average pooling and passed to a fully connected layer of 1024 units.

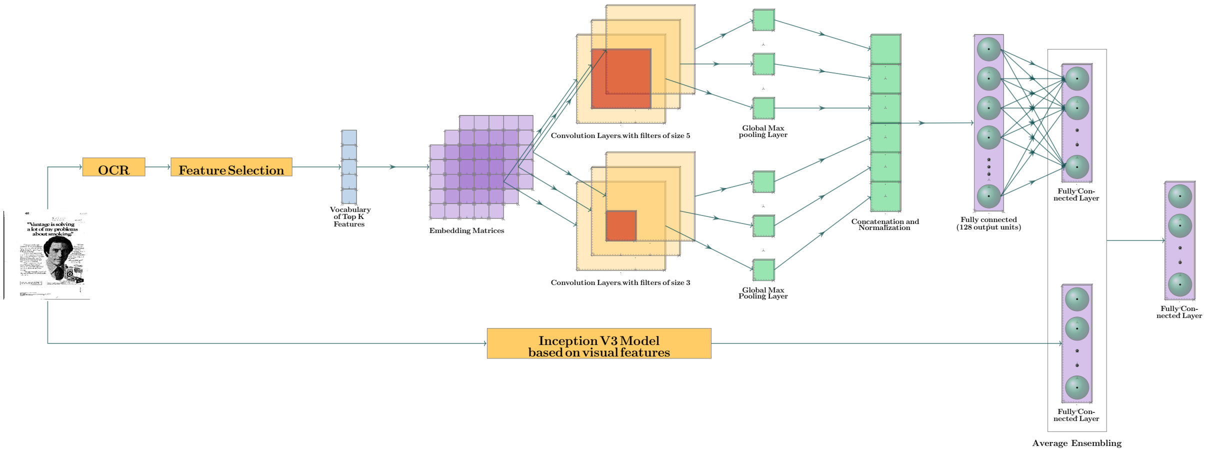

OCR
Feature Selection
Vocabulary of Top K Features
Embedding Matrices
Convolution Layers with filters of size 5
Global Max pooling Layer
Convolution Layers with filters of size 3
Global Max Pooling Layer
Concatenation and Normalization
Fully connected (128 output units)
Fully Con-nected Layer
Fully Con-nected Layer
Inception V3 Model based on visual features
Fully Con-nected Layer
Average Ensembling

Fig. 4: Two Stream Deep Network

Finally, the last fully connected layer categorizes the images using softmax activation.

B. Textual stream

As the input layer of textual stream requires textual content, therefore to acquire textual content from text document images, we used Tesseract OCR[1] which is based on LSTM and trained on large corpus of data. We analyse the output of OCR and find a lot of errors in the recognition, especially in the text document images of three classes (hand written, advertisement and notes). After extracting text, unnecessary symbols, characters, and stop words are removed. In order to select discriminative features, we have used filter based feature ranking algorithm namely Balanced Accuracy Measure (ACC2) [32]. Although state-of-the-art work by Noce et al. [8] has utilized weighting formula proposed by Penas et al [9], to rank the features considering their frequency in both positive, and negative class documents, however, $ACC2$ is more robust as it computes the absolute difference between the occurrence rate of particular feature in positive class documents (t_{pr}), and negative class documents (f_{pr}). In order to illustrate the effectiveness of $ACC2$ over weighting formula, lets consider a hypothetical scenario, where a feature is very frequent in only one document of the positive class but it does not appear at all in negative class documents. In this case, feature weighting formula [9] will assign higher score to this features despite having no occurrence in most of the positive class documents. However, as $ACC2$ utilize the document frequency, thus, it will assign a lower rank to such features. In this way, $ACC2$ assigns higher rank to those features which reveal higher document frequency for positive but lower document frequency for negative class. Mathematical expression of $ACC2$ is written as:

$$ACC2 = |t_{pr} - f_{pr}| \tag{1}$$

Moreover, although feature ranking metrices like NDM [33], and MMR [34] are advanced variations of $ACC2$, however both algorithms do not perform well for the text extracted by OCR. Both algorithms assign higher score to those features which have either $t_{pr} = 0$ or $f_{pr} = 0$), which happens due to the poor performance of OCR. As $ACC2$ does not consider min(t_{pr}, f_{pr}) or max(t_{pr}, f_{pr}), this is why it reveals better performance.

After applying pre-processing on the textual content extracted by OCR from document images, we rank the features of all classes using $ACC2$ feature ranking algorithm. Finally top 450 features from each class are selected and fed to multi-channel CNN. For sake of utilizing both syntactic and semantic features, multi-channel CNN is used. In our experimentation, OCR reveals a pretty poor performance, this is why only 25% of features got a hit in pre-trained word embeddings. In order to learn the embeddings of remaining 75% of features, embedding layers of two channels (multi-channel CNN) are randomly initialized. Multi-channel convolutional neural network consists of three channels where each channel starts with an embedding layer. Each embedding layer is followed by two convolutional layers with 16 filters of size 3 and 5 respectively. The features of all feature maps are first pooled using global average pooling layer and then concatenated and normalized using l2 normalization approach. After that, these normalized features are passed to a fully connected layer with 128 units. Finally, a fully connected layer with softmax activation acts as a classifier.

Graphical representation of Multi-channel model is shown in Figure 4

C. Embedding Visual and textual stream

While visual stream classifies the text document images into predefined classes using the spatial information, textual stream effectively utilizes the content of text document images. To reap the benefits of both streams, we embed visual and textual streams using an average ensembling approach in the proposed methodology. In proposed methodology, a class is assigned to unseen text document image that reveals highest weighted average computed by combining the predictions of both streams.

[1] https://github.com/tesseract-ocr/tesseract

V. Datasets

This section provides details of three datasets used in our experimentation. In order to assess the performance of two stream text document image classification approach, we use publicly available Tobacco-3482 dataset which contains 3482 text document images of ten diverse classes. Moreover, ImageNet and Ryerson Vision Lab Complex Document Information Processing (RVL-CDIP) datasets are utilized for transfer learning. ImageNet contains almost 1350000 images related to daily life things and RVL-CDIP has 400,000 text document images of 16 different classes. Standards split of ImageNet has 1.2 million training, 50,000 validation, and 100,000 testing samples, whereas RVL-CDIP contain 320,000 training 40,000 validation, 40,000 testing samples of text document images.

As Tobacco-3482 dataset is the subset of RVL-CDIP, this is why, overlapped images of tabbacco-3482 test set are removed from RVL-CDIP dataset.

Statistical information of extracted text for both RVL-CDIP, and Tobacco-3482 datasets are revealed in Table I.

Dataset	Documents	Classes	Features	Min Classes	Max Classes
RVL-CDIP	399743	16	4257214	24951	25000
Tobacco-3482	3482	10	130320	120	620

TABLE I: Statistics of RVL-CDIP, and Tobbacco-3482 datasets

VI. Experimental Setup

This section explains the experimental settings made for the proposed classification methodology which comprises of two phases. In visual stream, InceptionV3 is trained on RVL-CDIP dataset using the pre-trained weights of ImageNet. The model is trained for 45 epochs with the batch size of 32. Moreover, Adam optimizer with the learning rate 0.001, β_1 0.9, and β_2 of 0.999 is used, while categorical cross entropy is used as a loss function.

In textual stream, top 450 features ranked by balanced accuracy measure ($ACC2$) are selected and fed to the embedding layers of Multi-channel CNN. Embedding layer of one channel is initialized with pre-trained word vectors provided by fastText [2] as compared to other channels which are initialized randomly. The multi-channel CNN model is trained using mini batch of size 50 for 20 epochs. Furthermore, Categorical Cross Entropy is used as a loss function, whereas RMSprop is used as an optimizer with the learning rate of 0.001 and ∂ of 0.9.

VII. Results

This section compares the performance of proposed two stream text document image classification methodology with standalone InceptionV3 model based on transfer learning. In addition, it provides detailed comparison with state-of-the-art methods for text document image classification. On Tobacco-3482 dataset, confusion matrices of both proposed methodology and standalone InceptionV3 model are shown in Figure 5a and 5b respectively.

Author	Method	Accuracy
Chen et al. [2012] [19]	HVP-RP	40.3
Kang et al. [2014] [35]	CNN	65.35
afzal et al. [2015] [6]	Alexnet	77.6
Noce et al. [2016] [8]	CNN with combination of Textual and Visual features	79.8
Afzal et al. [2017] [36]	Resnet-50	91.3
Proposed Two Stream Approach	Average Ensembling of Textual and Visual features	95.8

TABLE II: Accuracy comparison of proposed two stream deep network with state-of-the-art document image classification methodologies

As the Figure 5a and 5b suggest, the proposed two stream classification methodology outperforms standalone Image based InceptionV3 approach with the figure of 4.5% on Tobacco-3482 dataset. The proposed methodology raises the performance of all classes except Email where it falls by the figure of just 1%. Furthermore, in case of highly overlapped classes (Letter, Memo, Report), standalone InceptionV3 approach only manages to classify 94% of Letter images accurately and wrongly categorize almost 5% of Letter images, whereas the proposed methodology classifies 98% of Letter images correctly and only 0.5% of Letter images are wrongly classified in Memo, and Report classes. Similar trend can be seen for Memo, and Report classes.

Table II compares the performance of the proposed two stream text document image classification methodology with the state-of-the-art methodologies. Amongst all state-of-the-art approaches, ResNet-50 managed to produce the highest figure of 91.3%. However, the proposed two stream network wit average ensembling methodology outperforms ResNet-50 with the figure of 2.78% on Tobacco-3482 dataset.

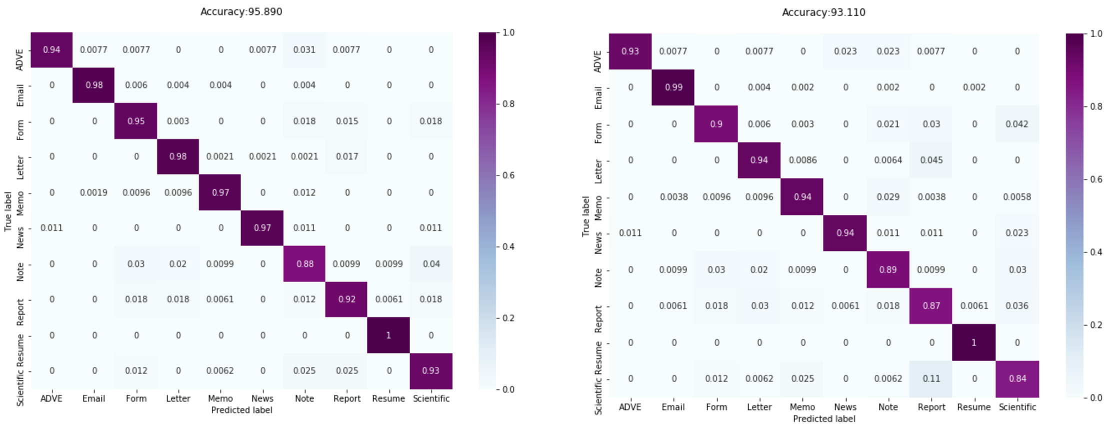
Furthermore, to analyze the integrity of vocabulary developed by ACC2 measure, we use same vocabulary of discriminative words (generated from Tobbaco dataset) on the test set of RVL dataset with only 9 classes of RVL dataset which directly overlaps with Tobacco dataset. As it can be seen from the Table III, the overall accuracy of the proposed Multi-Channel CNN has increased by almost 4% using ranked features as compared to full features. Classes like advertisement, form, letter, memo, news article and scientific publications have shown an increase in the accuracy by the margin of 2%, 4%, 8%, 9%, and 17% respectively. Moreover, the InceptionV3 model has achieved an accuracy of 93.2% independently, however with the combination of both InceptionV3 and Multi-Channel CNN, an overall accuracy of 96.4% is achieved.

Although Noce et al[8] used similar combination approach to produce the performance figure of 79.8%, however we manage to produce best performance figure of 95.8% because of better feature ranking algorithm and convolutional neural network architecture.

VIII. Conclusion

This paper presents a naive deep learning based approach for the task of text document image classification. Although, state-of-the-art deep learning based approaches are producing good results using transfer learning, however, these approaches fail

[2] https://fasttext.cc/docs/en/english-vectors.html



(a) Two Stream Methodology

(b) InceptionV3 Model

Fig. 5: (a) Confusion Matrix for Proposed Two stream Text Document Image Classification, (b) Confusion Matrix of InceptionV3 Model

Class Label	Accuracy of Multi-Channel CNN using Textual Features			Accuracy of CNN Model with Visual Features	Accuracy of two-stream approach with the combination of textual and Visual Features
	Full vocab	Discriminative set of Features	Δ (Full Vocab-DF)		
advertisement	85	87	2	93	96
email	93	84	-9	99	99
form	80	84	4	89	94
letter	82	90	8	92	96
memo	80	89	9	95	97
news article	80	89	9	94	97
resume	95	93	-2	95	98
scientific publication	74	91	17	92	96
scientific report	81	77	-4	90	95
Overall Accuracy	83.3	87.1	3.8	93.2	96.4

TABLE III: Performance of two stream classification methodology using discriminative features of tobacco-3482 dataset and evaluated on 9 classes of RVL-CDIP which are common in both datasets

to produce promising performance in case of those datasets where the classes highly overlap with each other. Thus, we propose a deep learning based hybrid methodology which utilizes both structural similarity and content of text document images in order to outshine existing state-of-the-art methodologies. In future, we will investigate various other ensembling approaches to raise the performance of text document image classification.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[2] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[5] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognition*, vol. 77, pp. 354–377, 2018.

[6] M. Z. Afzal, S. Capobianco, M. I. Malik, S. Marinai, T. M. Breuel, A. Dengel, and M. Liwicki, "Deepdocclassifier: Document classification with deep convolutional neural network," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2015, pp. 1111–1115.

[7] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[8] L. Noce, I. Gallo, A. Zamberletti, and A. Calefati, "Embedded textual content for document image classification with convolutional neural networks," in *Proceedings of the 2016 ACM Symposium on Document Engineering*. ACM, 2016, pp. 165–173.

[9] A. Peñas, F. Verdejo, J. Gonzalo *et al.*, "Corpus-based terminology extraction applied to information access," in *Proceedings of Corpus Linguistics*, vol. 2001. Citeseer, 2001, p. 458.

[10] M. N. Asim, M. Wasim, M. S. Ali, and A. Rehman, "Comparison of feature selection methods in text classification on highly skewed datasets," in *2017 First International Conference on Latest trends in Electrical Engineering and Computing Technologies (INTELLECT)*. IEEE, 2017, pp. 1–8.

[11] J. Kumar and D. Doermann, "Unsupervised classification of structurally similar document images," in *2013 12th International Conference on Document Analysis and Recognition*. IEEE, 2013, pp. 1225–1229.

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

[13] C. K. Shin and D. S. Doermann, "Document image retrieval based on layout structural similarity." in *IPCV*, 2006, pp. 606–612.

[14] J. Hu, R. Kashi, and G. Wilfong, "Comparison and classification of documents based on layout similarity," *Information Retrieval*, vol. 2, no. 2-3, pp. 227–243, 2000.

[15] P. Sarkar, "Learning image anchor templates for document classification and data extraction," in *2010 20th International Conference on Pattern Recognition*. IEEE, 2010, pp. 3428–3431.

[16] C. Shin, D. Doermann, and A. Rosenfeld, "Classification of document pages using structure-based features," *International Journal on Document Analysis and Recognition*, vol. 3, no. 4, pp. 232–247, 2001.

[17] J. Kumar, P. Ye, and D. Doermann, “Learning document structure for retrieval and classification,” in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE, 2012, pp. 1558–1561.

[18] ——, “Structural similarity for document image classification and retrieval,” *Pattern Recognition Letters*, vol. 43, pp. 119–126, 2014.

[19] S. Chen, Y. He, J. Sun, and S. Naoi, “Structured document classification by matching local salient features,” in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE, 2012, pp. 653–656.

[20] K. U. Reddy and V. Govindaraju, “Form classification,” in *Document Recognition and Retrieval XV*, vol. 6815. International Society for Optics and Photonics, 2008, p. 68150Y.

[21] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner *et al.*, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[22] L. Kang, J. Kumar, P. Ye, Y. Li, and D. Doermann, “Convolutional neural networks for document image classification,” in *2014 22nd International Conference on Pattern Recognition*. IEEE, 2014, pp. 3168–3172.

[23] A. W. Harley, A. Ufkes, and K. G. Derpanis, “Evaluation of deep convolutional nets for document image classification and retrieval,” in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2015, pp. 991–995.

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

[25] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[26] A. Kölsch, M. Z. Afzal, M. Ebbecke, and M. Liwicki, “Real-time document image classification using deep cnn and extreme learning machines,” in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 1. IEEE, 2017, pp. 1318–1323.

[27] Y. Zhang, S. Roller, and B. Wallace, “Mgnc-cnn: A simple approach to exploiting multiple word embeddings for sentence classification,” *arXiv preprint arXiv:1603.00968*, 2016.

[28] X. Zhang, J. Zhao, and Y. LeCun, “Character-level convolutional networks for text classification,” in *Advances in neural information processing systems*, 2015, pp. 649–657.

[29] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun, “Very deep convolutional networks for text classification,” *arXiv preprint arXiv:1606.01781*, 2016.

[30] S. Lai, L. Xu, K. Liu, and J. Zhao, “Recurrent convolutional neural networks for text classification,” in *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[31] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[32] G. Forman, “An extensive empirical study of feature selection metrics for text classification,” *Journal of machine learning research*, vol. 3, no. Mar, pp. 1289–1305, 2003.

[33] A. Rehman, K. Javed, and H. A. Babri, “Feature selection based on a normalized difference measure for text classification,” *Information Processing & Management*, vol. 53, no. 2, pp. 473–489, 2017.

[34] A. Rehman, K. Javed, H. A. Babri, and N. Asim, “Selection of the most relevant terms based on a max-min ratio metric for text classification,” *Expert Systems with Applications*, vol. 114, pp. 78–96, 2018.

[35] L. Kang, J. Kumar, P. Ye, Y. Li, and D. Doermann, “Convolutional neural networks for document image classification,” in *2014 22nd International Conference on Pattern Recognition*. IEEE, 2014, pp. 3168–3172.

[36] M. Z. Afzal, A. Kölsch, S. Ahmed, and M. Liwicki, “Cutting the error by half: Investigation of very deep cnn and advanced training strategies for document image classification,” in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 1. IEEE, 2017, pp. 883–888.