WordVIS: A Color Worth A Thousand Words

Umar Khan*, Saifullah, Stefan Agne, Andreas Dengel, Sheraz Ahmed

DFKI, Kaiserslautern

Abstract. Document classification is considered a critical element in automated document processing systems. In recent years multi-modal approaches have become increasingly popular for document classification. Despite their improvements, these approaches are underutilized in the industry due to their requirement for a tremendous volume of training data and extensive computational power. In this paper, we attempt to address these issues by embedding textual features directly into the visual space, allowing lightweight image-based classifiers to achieve state-of-the-art results using small-scale datasets in document classification. To evaluate the efficacy of the visual features generated from our approach on limited data, we tested on the standard dataset Tobacco-3482. Our experiments show a tremendous improvement in image-based classifiers, achieving an improvement of 4.64% using ResNet50 with no document pre-training. It also sets a new record for the best accuracy of the Tobacco-3482 dataset with a score of 91.14% using the image-based DocXClassifier with no document pre-training. The simplicity of the approach, its resource requirements, and subsequent results provide a good prospect for its use in industrial use cases.

Keywords: document image classification · document classification · image processing · data pre-processing · visual embeddings · textual embeddings · deep learning.

1 Introduction

Documents play an integral role in modern business communication and record keeping. As a result, there is a growing interest today in the development of automated document processing pipelines in business workflows [1, 22, 27, 29]. Document classification is a common starting point for many of these pipelines. Early classification of documents not only facilitates filtering, searching, and retrieval, but also enhances downstream performance [1, 10, 18]. For example, if it is possible to categorize a specific class with a high degree of confidence, an efficient information extraction module may be developed particularly for that class, thereby enhancing the pipeline's overall efficiency. Due to its fundamental

importance, document classification has been extensively studied over the past few decades [1, 7, 11, 21, 27] and has been widely adopted by the industry.

Deep learning has been extensively studied in recent years in the context of document classification, resulting in a variety of approaches both in the image domain and in the multimodal domain. As a result, the task of building a high-performance document classifier is no longer considered arduous. As long as sufficient computing resources and training data are available, there are numerous state-of-the-art approaches [4, 15, 24, 25, 27] that can be directly applied to both large and small datasets to produce exceptional results in document classification. Especially interesting are the recent multimodal approaches which use both visual and textual features to perform the classification task and are particularly successful at countering the problem of high inter-class similarity and intra-class variance commonly found in documents [16, 18]. However, there are several challenges involved with these approaches from a deployment perspective. Firstly, many of these approaches involve multiple streams of networks [4, 23] or large multimodal transformer networks [24, 27], which greatly increase the computational load. Additionally, such models are particularly challenging to train, since they require self-supervised learning across millions of data points [27, 28]. This can be especially problematic for small businesses, which have limited computing resources or training data, making it difficult to deploy such approaches in a practical manner. In addition to these issues, most existing multimodal approaches require feeding the textual and layout information directly to the model [4, 27, 28], which may require overhauling an existing document processing system. Finally, such multimodal techniques can also be difficult to extend to new languages and require additional training data for each target language.

In this paper, we attempt to counter the aforementioned issues in multimodal approaches and present a lightweight approach for document classification that utilizes both visual and textual features of a document image without the need for any kind of self-supervised pretraining. Contrary to most existing multimodal approaches, we embed the textual semantic features and context directly into the visual space of a document by assigning an RGB color to each word in accordance with the similarity of different letters. In this manner, our approach not only allows existing image-based classifiers to directly exploit the textual cues of a document but also substantially reduces the performance overhead associated with the processing of multi-modal data. In addition, it requires no additional pretraining to learn the textual embeddings as in the case of typical multi-modal approaches, making it particularly suitable for small datasets. Due to the simplicity of our approach, not only can it easily be integrated into existing CNN-based document classification pipelines but can also be directly extended for any new languages without the need for additional language-specific data, as is the case with most transformer-based approaches. The overall contributions of our paper can be summarized as follows:

- We present a novel approach for embedding textual semantic and contextual features in the visual space of a document. This enables training high-performing document classifiers in data-scarce settings.

– For the ablation study, we evaluate our approach with multiple CNN-based architectures on a small-scale document benchmark dataset Tobacco-3482 and show that our approach results in consistent performance improvements ranging from $3-5\%$ simply by training the models on the dataset.
– WordVis was also able to improve the performance of the state-of-the-art DocXClassifier-B, ultimately resulting in a new best-record accuracy score of 91.14% without the use of RVL-CDIP document pre-training, which means enabling the development of sustainable classifiers without extensive training, especially in data-scarce situations.

2 Related Work

2.1 Document Image Classification

The field of document image classification has evolved considerably over the past few decades. The early works in the field were primarily based on structural similarities between documents [26], feature matching [21], or their combination [9]. In a different direction, classical machine learning-based approaches such as K-Nearest Neighbors (KNN) [6], Decision Trees [7] and Hidden Markov Models (HMM) [14] have also been explored for this task. A detailed overview of these approaches can be found in the survey paper by Chen *et al.* [8].

It was not long after the seminal work by Krizhevsky *et al.* [20] in which the popular AlexNet architecture was introduced for natural image classification, that a range of deep-learning based document classification systems were introduced. Kang *et al.* [18] were the first to demonstrate the effectiveness of a neural network for document classification, which was significantly more successful compared to classical approaches. Later, Afzal investigated the use of much deeper CNNs for the classification of documents, as well as the advantages of transfer learning for document classification. Since then, this trend has continued steadily with the use of newer versions and variations of CNNs for the classification of documents [15, 25]. As CNN-based approaches have started to reach diminishing returns for this task, there has also been a growing interest in multi-modal approaches that combine the image, layout, and textual information of the document to perform the classification task. These techniques are generally implemented either in a multi-stream fashion which uses separate streams for visual, textual or other layout features [4,23] or based on multimodal transformer architectures [27,28] that are trained in a self-supervised manner on millions of training samples.

2.2 Visual Encoders for Textual Information

In this section, we review some previous works for encoding textual semantics or contextual information into visual space in document analysis. Anoop *et al.* [19] proposed Chargrid, a grid-based color coding scheme for document images where the bounding box region of each character is colorized on a separate image

mask based on a predetermined encoding scheme. The original image and its corresponding colorized image mask were then used to train a neural network for the task of key information extraction (KIE). To introduce more contextual and semantic information into the textual encoding, Timo *et al.* [12], introduced BERTGrid for the KIE task in which instead of colorizing the document in RGB space, instead they concatenate the BERT-based [13] textual embeddings with the image. In particular, for each word, its corresponding BERT embedding is concatenated on top of the image within its bounding box. While this approach provided richer contextual information in comparison to Chargrid, it came with huge computational costs and increased dimensionality of the image input. Lin *et al.* [22] presented a similar approach to BERTGrid for KIE task but instead introduced the concatenation step of textual embedding into visual space at an intermediate feature map of a Convolutional Neural Network (CNN) instead of concatenating it with the input image. In a slightly different direction, Saman *et al.* [3] used a simple color encoding scheme with the idea of distinguishing between numbers and alphabets to colorize the document images for the task of table detection in document images. As mentioned above, while some work has been done to encode textual information into visual space, it has not been explored in the context of document classification.

3 WordVis: The Proposed Approach

WordVIS is a novel pre-processing method that generates enhanced document representations using OCR data. It generates enriched visual representation by encoding text to colors using a score lookup table and applies color masks on words using the original document image. The scoring mechanism was inspired by the fundamental concept of string metric calculation from information theory called "Levenshtein Distance". We leverage the core concept of how the distance is measured in strings. In this section, we describe the process of generating these document representations using WordVIS.

3.1 Score Assignment

Considering all the characters N_c in the language characters space, the limit on supported characters is $\lim_{1\to\infty} N_c$. However, each character in the character space will be assigned a score C_s representing the weightage of that character. The limit for the weightage of a character can be defined as $\lim_{0\to255} C_s$, as the maximum score possible for the $\sum_{n=0}^{n=\infty} C_s$ is 255 per channel as that's the maximum value possible in the individual channels of the RGB representation of the color. Given the target dataset is Tobacco-3482, all references in this research publication will point to the use of English language characters. However, it is to be noted that the method itself is language agnostic.

Purpose The score assignment process is about assigning weightage C_s to individual characters. This allows the users of the system to leverage weights on

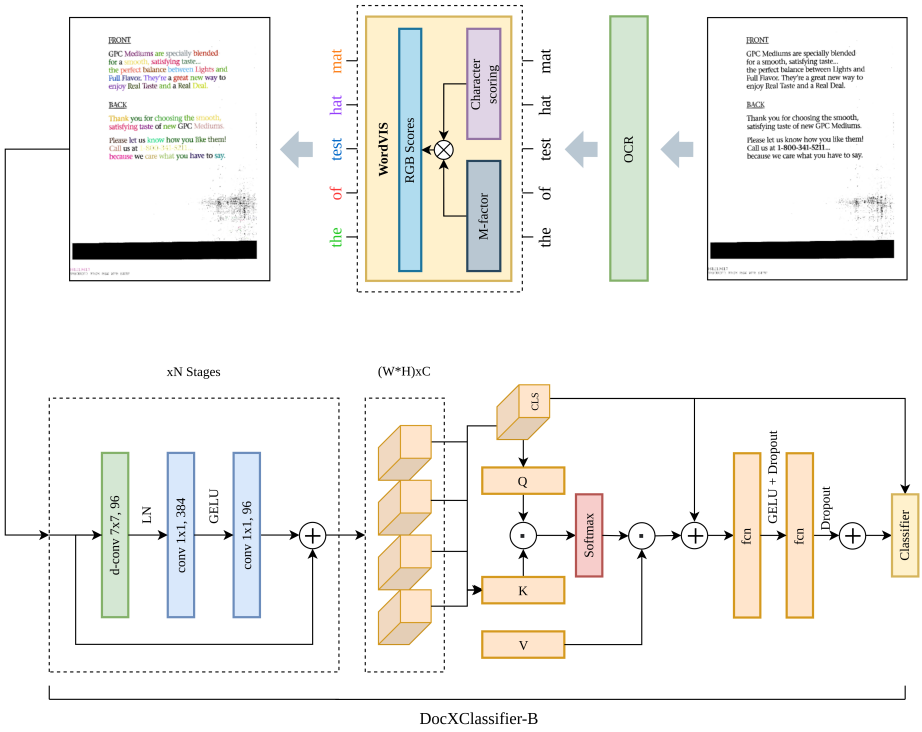


Fig. 1: WordVIS as a pre-processing step for existing document classification models. Input document images are first passed through an OCR system to extract textual information. The textual features are then encoded within the visual space of the images using our approach. Finally, the pre-processed images are fed into a document classification model. Shown here is the DocXClassifier-B model.

individual characters depending on their semantic understanding of their dataset and the need for it. This weightage score forms the basis for the RGB number of the word. If the weights are well distributed on the maximum aforementioned range would give us a unique color for all possible words, as the maximum possible combinations using this scheme are 16777216 which is a far greater number than all possible words in the English language.

Distribution Scheme The method is flexible to the color score distribution scheme adopted by the users in accordance with the desired output color range. As the number of maximum possible scores per individual character is $NC_s = 3$ one each for the R,G and B channels, the coloring scheme can be based on a single character having a score in all three channels, two out of three or a single channel.

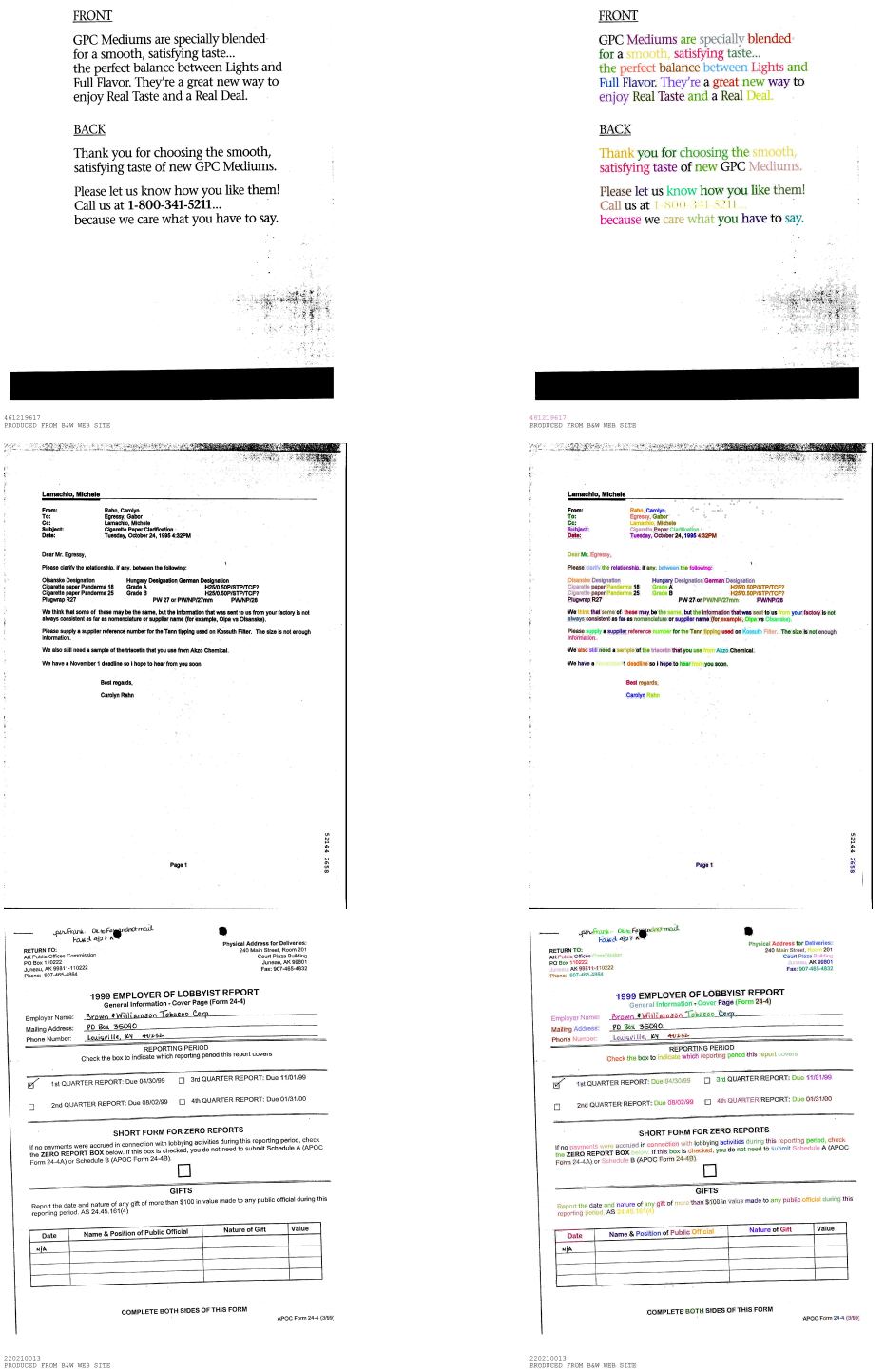


Fig. 2: WordVIS samples of different classes produced. We can see that most of the textual data is masked with colors without changing non-textual elements of the document images.

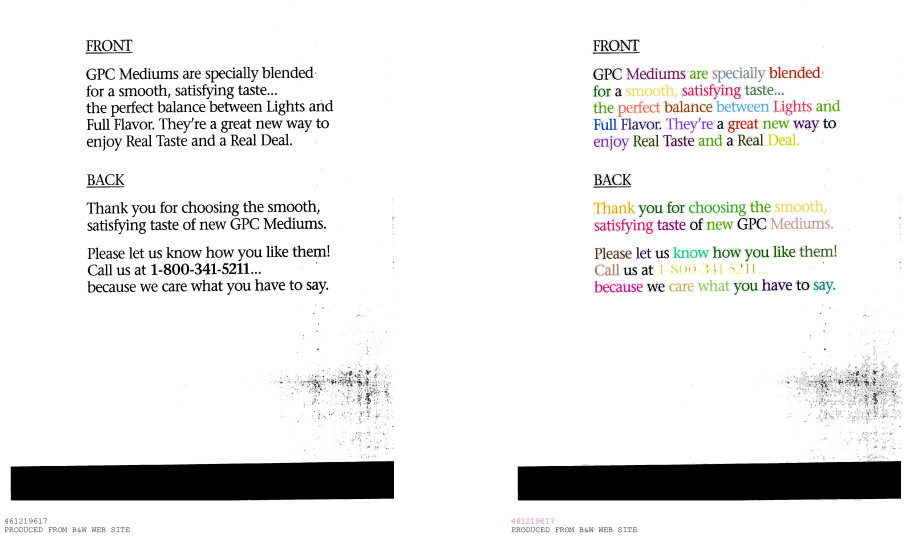


Fig. 3: In WordVIS colorized document we can see that the colorization adapts a pattern of stop words or words more similar to stop words adapting a green color, whereas the more lengthier and distinct words adapting more distinct colors.

Experiment Scores For our experiments, there are 26 English language characters and 10 digits taking our $N_c = 36$. The assignment operation can be both generic and very targeted towards a specific dataset leveraging the semantic understanding of that dataset. However, for a more fair comparison of the standard dataset, we have chosen to assign more generic scores to the characters of the English language without the use of special characters or semantic understanding of the test dataset. Our range of weights for our experiments is $\lim_{1 \to 9} C_s$ however, we believe they can be used to further enhance results with little insights into the dataset. Our experimentation scoring scheme is the division of all available characters between Red, Green, and Blue channels with each channel having unique 12 characters. Our score assignment per character is generic in ascending order for all characters and without the use of special characters for avoiding any semantic encoding using the dataset knowledge.

3.2 Multiplying Factor M

In order to bring a weight on the length of the word that occurs in the text corpus, we introduced a multiplying factor M_f based on the length of the word. The Multiplying Factor is what ensures that minimally assigned scores in the character space are elevated enough to translate into sharp colors based on the word length. M can be both heuristically assigned according to insights into the data as well as set more generically as we did with our multiplying factor being derived from the word character length itself. In our case, every word will

have this factor differently assigned from its character length, and hence why lengthier words will translate to stronger colors than shorter words. This value in combination with the C_s can be used to build a tighter range of the $\sum C_s$.

3.3 Calculating RGB Values

The formula that is used for converting the individual scores to RGB is given below in Fig. 1 and Fig. 2

$$R = \sum_{C=a}^{C=i} C_s * M_f, G = \sum_{C=j}^{C=r} C_s * M_f, B = \sum_{C=s}^{C=z} C_s * M_f \tag{1}$$

$$RGB_{\text{color}} = (R, G, B) \tag{2}$$

Whereas C_s is the score of the individual character and M_f is the multiplying factor derived from $Word_{\text{length}}$.

Example The above algorithm can be more fairly understood from the following calculation example: Given the word "deep", the RGB score according to the above example would be

$Word_{\text{length}} = 4 = \text{M}$
$R_{\text{score}} = ((\text{d=3}) * 4) + ((\text{e=5}) * 4) + ((\text{e=5}) * 4) = 52$
$G_{\text{score}} = ((\text{p=7}) * 4) = 28$
$B_{\text{score}} = ((0) * 4) = 0$

$RGB_{\text{score}} = (52, 28, 0)$

This example highlights a property of WordVIS, where small errors in OCR are limited by the constraints put on the scoring mechanism leading to more consistent colors even in case of errors, to further continue the previous example: If the given word "deep" was incorrectly OCR'ed as "deeq", the resulting change would be minimal as shown in the calculation below. $Word_{\text{length}} = 4 = M$
$R_{\text{score}} = ((d = 3) * 4) + ((e = 5) * 4) + ((e = 5) * 4) = 52$
$G_{\text{score}} = ((p = 8) * 4) = 32$
$B_{\text{score}} = ((0) * 4) = 0$

$RGB_{\text{color}} = (52, 32, 0)$

3.4 Output and Sample Analysis

The following samples shows the resulting documents of several classes from the WordVIS visual embedding on Tobacco-3482 documents:

Table 1: A comparison of the classification accuracy of different approaches on the Tobacco3482 datasets

Model	Inference Time [ms]	# of Parameters	Tobacco3482 (ImageNet pre-training)
AlexNet (Afzal *et al.*, 2017 [2])	1.1	57M	75.73%
GoogleNet (Afzal *et al.*, 2017 [2])	1.2	5.6M	72.98%
ResNet-50 (Afzal *et al.*, 2017 [2])	1.1	23.5M	67.93%
VGG-16 (Afzal *et al.*, 2017 [2])	1.3	134M	77.52%
EfficientNet (Ferrando *et al.*, 2020 [15])	2.3	17.6M	85.99%
EfficientNet + BERT (Ferrando *et al.*, 2020 [15])	-	127.6M	89.47%
MobileNetV2+Text (Audebert *et al.*, 2019 [5])	-	-	87.80%
EfficientNet+BERT (Kanchi *et al.*, 2022 [17])	-	197M	90.3%
DocXClassifier-B/384	6.53	95.4M	88.42%
DocXClassifier-L/384	10.0	204M	88.43%
DocXClassifier-XL/384	16.1	356M	90.14%
WordVIS DocXClassifier-B/384 (OURS)	**6.53**	**95.4M**	**91.14%**

Fig. 2 shows that the images are pre-processed utilizing varying degrees of colors using the scoring mechanism. Through further inspection, we notice that most of the stop words due to them containing the same letters repeating, contain the same colors and non-stop words tend to take more different colors than the stop words. This property can be observed in all classes, from our Fig. 3 in class ADVE it can be observed that words like "you, are, the, and, for, new, what" all have green hues, whereas words such as "satisfying, blended, taste, medium, smooth" all take sharper and different tones of colors.

4 Experiments and Results

To evaluate the efficacy of our proposed method we trained several architectures on the standard dataset of Tobacco-3482. The smaller dataset was used in order to test for the performance gain on small limited amounts of datasets. We trained all these architectures on two different versions of the dataset.

- Standard Tobaccoo-3482 dataset
- WordVIS Colorized Tobaccoo-3482 dataset

4.1 Dataset

The dataset consists of a total of 3482 documents split across 10 classes. The splits were generated using random $80\% - 20\%$ train-test split resulting in 700 test documents. The validation split was 10% of the training data resulting in 2504, 278, 700 samples in Train, Validation, and Test respectively. Moreover, it is to be noted that these experiments were performed multiple times with the given splits to eliminate the standard deviation as the cause for higher accuracy.

4.2 Training Details

For hyperparameters consistency, we ensure the same hyperparameters for both with and without the use of WordVIS to eliminate any issues that could be caused by differences in hyper parameters. For training the DocXClassifier-B the original hyperparameters used by authors of DocXClassifier Saifullah et al [25] were used. All the remaining architectures of ResNet50, ResNet101, densenet121, EfficientNetV2 were trained with the same hyperparameters. For the optimizer Stochastic Gradient Decent (SGD) was used with a learning rate of: 0.5, weight decay was set to 1.0e-8, and batch size of 64. All the architectures were trained for 300 epochs, however, the convergence occurred for all before the 80 epoch mark. ferrandodc

4.3 Evaluation

For comparative analysis using a state-of-the-art document classifier on Tobacco-3482 dataset, we took the DocXClassifier-B, the smallest in terms of the number of parameters in the state-of-the-art category, and tested WordVIS. As seen in Table 1, our method WordVIS improves on the base model drastically, setting a new record score on the Tabaccoo-3482 dataset of 91.14%. It is to be noted that not only does it improve drastically on the base class but also outperforms the previous state-of-the-art DocXClassifier-XL while using 73.2% fewer parameters and reducing the inference time by 60%.

4.4 Ablation Study

To further test our method, we also trained several CNN variant architectures. As we can see from Table 2 that the WordVIS method outperformed the no pre-processing on all architectures.

4.5 Qualitative Analysis

In order to analyze and verify the improvements brought in by WordVIS, we generated activation HeatMaps for test samples. As seen from Fig. 4 we can see

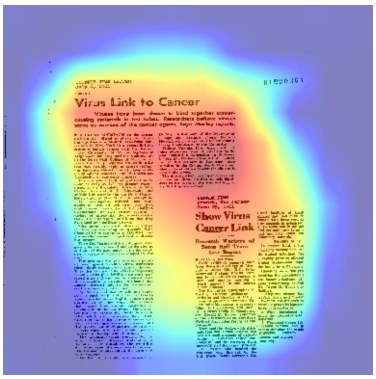
Table 2: Ablation study: A comparison of different architectures with and without WordVIS

Model	Base	WordVIS
ResNet50	67.9%	**72.5%**
ResNet101	79.1%	**82.0%**
densenet121	77.6%	**80.3%**
EfficientNetV2	74.1%	**76.3%**

ADVE-WordVis

ADVE-w/o WordVis

News-WordVis

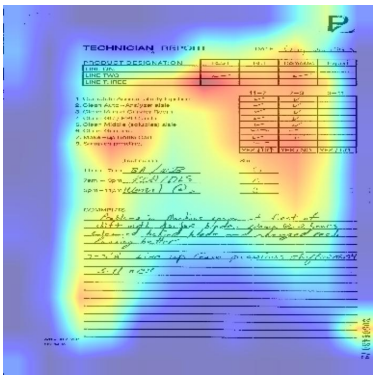


News-w/o WordVis

Email-WordVis

Email-w/o WordVis

Form-WordVis



Form-w/o WordVis

Fig. 4: Sample heatmaps generated using DocXClassifier-B with WordVIS (Left) and Without WordVIS (Right)

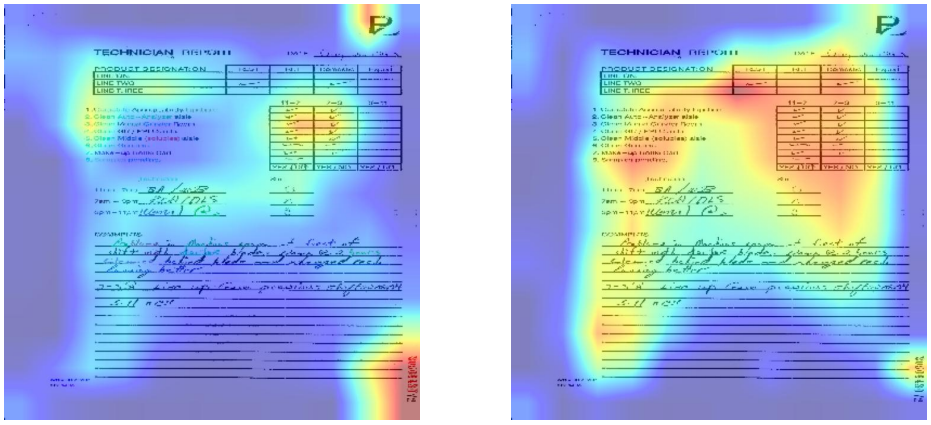


Fig. 5: Heatmaps Form: Form heatmaps also gives us clues into how the WordVIS (Left) trained network focuses on boxes and the content inside the boxes as opposed to the Base (Right).

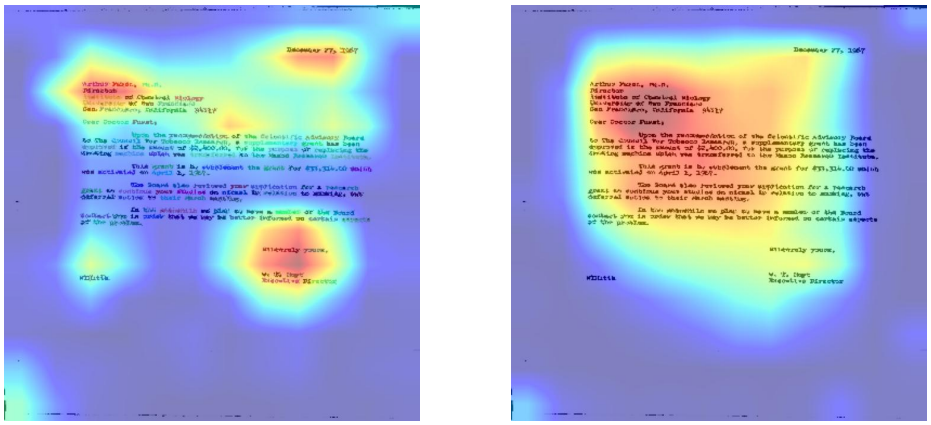


Fig. 6: Heatmaps of Letters shows us more more in depth on how the WordVIS method (Left) helps the network focus in on specific textual elements unique to the letter, whereas the training without the WordVIS focuses on the image as a whole

that the WordVIS (Top Row) shows a more refined focus toward specific text areas instead of a generalized focus on the entire image space, suggesting that the network learns to follow the text. In the example of Form Fig. 5 we can see that the WordVIS-trained DocXClassifier-B focuses on the specific boxed area and the activation is quite well tightly bounded as opposed to the base-trained DocXClassifier-B. Additionally, going through samples of Letters as depicted in Fig, 6 we observe a pattern of the WordVIS-trained network focusing on letter

headers and enclosers specifically rather than the entire body, in essence, closer to how humans identify letters as opposed to the Base trained network where the activation is spread on the entire image.

5 Conclusion and Future Work

In this research publication, we presented WordVIS, a novel pre-processing approach for utilizing text in Document Classification on smaller datasets. The heatmap analysis conducted in this research proves that our approach successfully embedded textual data in visual space, leading to the elimination of the textual embedding overhead used in multi-modal approaches. The research also successfully proves that image-based networks can be used to learn contextual text data by enhancing the textual data in document images. This approach not only reduces the required training data and compute overhead but also is able to efficiently leverage small amounts of data to outperform larger networks. Furthermore, the heatmap analysis also proves that such textual coloring techniques not only improve the quantitative accuracy but also drastically improve the quantitative results of the document classifiers. Our goal with this research was to enable businesses with limited training data and compute resources to be able to use and leverage document classifiers for their business use cases. We feel this is a step in the right direction, however, more research needs to be conducted to expand our work in order to improve accuracy on relatively simpler Deep Learning networks and reduce the architectural complexity requirements for more tolerable results.

References

1. Afzal, M.Z., Kolsch, A., Ahmed, S., Liwicki, M.: Cutting the Error by Half: Investigation of Very Deep CNN and Advanced Training Strategies for Document Image Classification. Proc. Int. Conf. Doc. Anal. Recognition, ICDAR **1**, 883–888 (2017)
2. Afzal, M.Z., Kölsch, A., Ahmed, S., Liwicki, M.: Cutting the error by half: Investigation of very deep CNN and advanced training strategies for document image classification. CoRR **abs/1704.03557** (2017), `http://arxiv.org/abs/1704.03557`
3. Arif, S., Shafait, F.: Table detection in document images using foreground and background features. In: Digital Image Computing: Techniques and Applications 2018. pp. 1–8 (2018)
4. Asim, M.N., Khan, M.U.G., Malik, M.I., Razzaque, K., Dengel, A., Ahmed, S.: Two stream deep network for document image classification. Proc. Int. Conf. Doc. Anal. Recognition, ICDAR pp. 1410–1416 (2019)
5. Audebert, N., Herold, C., Slimani, K., Vidal, C.: Multimodal deep networks for text and image-based document classification. CoRR **abs/1907.06370** (2019), `http://arxiv.org/abs/1907.06370`
6. Baldi, S., Marinai, S., Soda, G.: Using tree-grammars for training set expansion in page classification. Proc. Int. Conf. Doc. Anal. Recognition, ICDAR **2003-Janua**(Icdar), 829–833 (2003)

7. Cesarini, F., Lastri, M., Marinai, S., Soda, G.: Encoding of modified x-y trees for document classification. pp. 1131–1136 (01 2001). https://doi.org/10.1109/ICDAR.2001.953962
8. Chen, N., Blostein, D.: A survey of document image classification: Problem statement, classifier architecture and performance evaluation. Int. J. Doc. Anal. Recognit. **10**(1), 1–16 (2007)
9. Collins-Thompson, K., Nickolov, R.: A Clustering-Based Algorithm for Automatic Document Separation. Proc. SIGIR 2002 Work. Inf. Retr. OCR From Convert. Content to Grasping Mean. (September 2002) (2002)
10. Das, A., Roy, S., Bhattacharya, U., Parui, S.K.: Document Image Classification with Intra-Domain Transfer Learning and Stacked Generalization of Deep Convolutional Neural Networks. Proc. - Int. Conf. Pattern Recognit. **2018-Augus**, 3180–3185 (2018)
11. Dengel, A., Dubiel, F.: Clustering and classification of document structure-a machine learning approach. Proc. Int. Conf. Doc. Anal. Recognition, ICDAR **2**, 587–591 (1995)
12. Denk, T.I., Reisswig, C.: Bertgrid: Contextualized embedding for 2d document representation and understanding. CoRR **abs/1909.04948** (2019), `http://arxiv.org/abs/1909.04948`
13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf. vol. 1, pp. 4171–4186. Association for Computational Linguistics (ACL) (2019)
14. Diligenti, M., Frasconi, P., Gori, M.: Hidden tree Markov models for document image classification. IEEE Trans. Pattern Anal. Mach. Intell. **25**(4), 519–523 (2003)
15. Ferrando, J., Domínguez, J.L., Torres, J., García, R., García, D., Garrido, D., Cortada, J., Valero, M.: Improving accuracy and speeding up document image classification through parallel systems. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) **12138 LNCS**, 387–400 (2020)
16. Harley, A.W., Ufkes, A., Derpanis, K.G.: Evaluation of deep convolutional nets for document image classification and retrieval. Proc. Int. Conf. Doc. Anal. Recognition, ICDAR **2015-Novem**, 991–995 (2015)
17. Kanchi, S., Pagani, A., Mokayed, H., Liwicki, M., Stricker, D., Afzal, M.Z.: Emmdocclassifier: Efficient multimodal document image classifier for scarce data. Applied Sciences **12**(3) (2022). https://doi.org/10.3390/app12031457, `https://www.mdpi.com/2076-3417/12/3/1457`
18. Kang, L., Kumar, J., Ye, P., Li, Y., Doermann, D.: Convolutional neural networks for document image classification. Proc. - Int. Conf. Pattern Recognit. pp. 3168–3172 (2014)
19. Katti, A.R., Reisswig, C., Guder, C., Brarda, S., Bickel, S., Höhne, J., Faddoul, J.B.: Chargrid: Towards understanding 2d documents. CoRR **abs/1809.08799** (2018), `http://arxiv.org/abs/1809.08799`
20. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems. vol. 25 (2012)
21. Kumar, J., Ye, P., Doermann, D.: Structural similarity for document image classification and retrieval. Pattern Recognit. Lett. **43**(1), 119–126 (2014)

22. Lin, W., Gao, Q., Sun, L., Zhong, Z., Hu, K., Ren, Q., Huo, Q.: Vibertgrid: A jointly trained multi-modal 2d document representation for key information extraction from documents (2021). https://doi.org/10.48550/ARXIV.2105.11672, https://arxiv.org/abs/2105.11672
23. Noce, L., Gallo, I., Zamberletti, A., Calefati, A.: Embedded textual content for document image classification with convolutional neural networks. In: Proceedings of the 2016 ACM Symposium on Document Engineering. p. 165–173. DocEng '16, Association for Computing Machinery, New York, NY, USA (2016)
24. Powalski, R., Borchmann, Ł., Jurkiewicz, D., Dwojak, T., Pietruszka, M., Pałka, G.: Going Full-TILT Boogie on Document Understanding with Text-Image-Layout Transformer. In: Doc. Anal. Recognit. – ICDAR 2021. vol. 12822 LNCS, pp. 732–747 (2021)
25. Saifullah, Agne, S., Dengel, A., Ahmed, S.: Docxclassifier: High performance explainable deep network for document image classification (Mar 2022). https://doi.org/10.36227/techrxiv.19310489.v2
26. Shin, Christian and Doermann, D.: Document Image Retrieval Based on Layout Structural Similarity. Proc. 2006 Int. Conf. Image Process. Comput. Vision, Pattern Recognit. **2**, 606–612 (2016)
27. Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W., Zhang, M., Zhou, L.: LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding. pp. 2579–2591. Association for Computational Linguistics (ACL) (dec 2021)
28. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: LayoutLM: Pre-training of Text and Layout for Document Image Understanding. Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. **20**, 1192–1200 (2020)
29. Yu, W., Lu, N., Qi, X., Gong, P., Xiao, R.: Pick: Processing key information extraction from documents using improved graph learning-convolutional networks (2020). https://doi.org/10.48550/ARXIV.2004.07464, https://arxiv.org/abs/2004.07464