Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2022 doi:10.20944/preprints202202.0058.v1

Are Deep Models Robust against Real Distortions? A Case Study on Document Image Classification

Saifullah*†, Shoaib Ahmed Siddiqui*, Stefan Agne*‡, Andreas Dengel*†, Sheraz Ahmed*‡
*German Research Center for Artificial Intelligence (DFKI) 67663 Kaiserslautern, Germany
Email: {saifullah.saifullah, shoaib_ahmed.siddiqui,stefan.agne,andreas.dengel,sheraz.ahmed}@dfki.de
†TU Kaiserslautern, 67663 Kaiserslautern, Germany
‡DeepReader GmbH, 67663 Kaiserlautern, Germany

Abstract*—Deep neural networks have been extensively researched in the field of document image classification to improve classification performance and have shown excellent results. However, there is little research in this area that addresses the question of how well these models would perform in a real-world environment, where the data the models are confronted with often exhibits various types of noise or distortion. In this work, we present two separate benchmark datasets, namely RVL-CDIP-D and Tobacco3482-D, to evaluate the robustness of existing state-of-the-art document image classifiers to different types of data distortions that are commonly encountered in the real world. The proposed benchmarks are generated by inserting 21 different types of data distortions with varying severity levels into the well-known document datasets RVL-CDIP and Tobacco3482, respectively, which are then used to quantitatively evaluate the impact of the different distortion types on the performance of latest document image classifiers. In doing so, we show that while the higher accuracy models also exhibit relatively higher robustness, they still severely underperform on some specific distortions, with their classification accuracies dropping from ∼90% to as low as ∼40% in some cases. We also show that some of these high accuracy models perform even worse than the baseline AlexNet model in the presence of distortions, with the relative decline in their accuracy sometimes reaching as high as 300-450% that of AlexNet. The proposed robustness benchmarks are made available to the community and may aid future research in this area.

***Index Terms*—Document Image Classification, Corruption Robustness, Robustness to Distortions, Model Robustness**

I. Introduction

Deep learning has been applied to a number of challenging problems in recent years and has shown exceptional performance, particularly in computer vision [1], [2], [3], [4], natural language processing [5], and speech recognition [6]. However, recent studies [7], [8], [9] have shown that most modern machine learning models, while exceptionally powerful, are also quite fragile and are unable to robustly generalize over shifts in the data distribution. The problem is that these models rely heavily on the training data to be able to faithfully represent the data that will be encountered during deployment. However, in the real world, the data can be naturally corrupted [9], the data distribution can change over time [10], and the models are often confronted with new scenarios [11]. Another problem commonly found in modern machine learning models is that they are not able to identify when they are likely to be wrong, nor can they estimate how uncertain they are in their predictions [12]. Instead, they produce highly confident predictions during training and when an out-of-distribution data is encountered, the models begin to make erroneous yet confident predictions [13], a behavior that raises serious concerns about the reliability of these models.

Document image classification is one of the areas that has seen great success with the advent of convolutional neural networks [14], [15], [16] and, more recently, the transformers networks [17], [18]. This is also one of the areas where the distribution of training data and deployment data can be very different for various reasons. For example, the underlying templates for different classes of documents tend to vary significantly across organizations, making it difficult for deep networks to generalize well between different data distributions of the same class of documents. Another reason is that many of the documents found in the real-world are either altered (e.g., a form or questionnaire filled out by different users), naturally distorted, or damaged over time. Finally, with the increasing prevalence of smartphones in our society, many business processes have allowed the use of smartphones to scan documents, which can introduce additional types of distortions into the data. For example, images captured with mobile phones may exhibit various types of blur, noise, or data transformations.

Many deep networks have been proposed over the past few years that have reached the state of the art performance in both image-based [19] and multimodal document classification [18]. However, there is very little research that analyzes the robustness of these models on data outside of the distributions on which these models were trained. This analysis is particularly important as newer techniques are moving towards multi-modal approaches that combine visual and textual features to produce document representations, and are therefore highly dependent on how accurately the OCR (optical-character-recognition) software is able to extract the textual data from a document image. We believe that multimodal approaches may be severely affected by the introduction of distortions in the document images, as these could also lead to poor OCR results, which in turn could compromise the accuracy of these models. In this study, however, we limit ourselves to the uni-modal classification. To find out how well the existing approaches perform against different types of data distortions, we follow the approach of [9] and propose two robustness

© 2022 by the author(s). Distributed under a Creative Commons CC BY license.

benchmark datasets for document image classification, namely RVL-CDIP-D and Tobacco3482-D, generated by introducing 21 different types of document data distortions to the two well-known document classification benchmark datasets RVL-CDIP [14] and Tobacco3482, respectively. To summarize, the contributions of this work are as two-fold:

- We propose two benchmark datasets, RVL-CDIP-D and Tobacco3482-D, which can be used to evaluate the robustness of document image classification models to common real-world distortions.
- We use the proposed benchmarks to evaluate the performance of existing state-of-the-art document image classification models and present a thorough comparative analysis of their performance against common distortions.

II. Related Work

A. Document Image Classification

The field of document image classification has been extensively researched over the years. Early attempts in this area focused mainly on one of three possibilities; exploiting layout or structural similarity between documents [20], [21], feature extraction and matching [22], [23], or a combination of both [24]. Classic machine learning approaches such as K-Nearest Neighbours [25], and Hidden Markov Models [26] have also been proposed in the past. For a detailed overview of the older techniques, we refer the readers to a related survey [27].

With the advent of deep learning, the use of Convolutional Neural Networks (CNNs) has become increasingly popular in this field. Kumar et al. (2014) [28] was the first to demonstrate that CNNs can significantly outperform classical approaches even with a simple shallow network. The breakthrough came when Afzal et al. (2015) [16] and Harley et al. [14] (2015) showed that transfer learning from ImageNet pre-trained networks can extraordinarily boost the performance of existing CNNs. Afzal et al. (2017) [15] further improved performance by combining the transfer-learning approach with much deeper networks. Das et al. (2018) [29] proposed ensembles of multiple region-based classifier models to slightly improve the performance. Ferrando et al. (2020) [19] studied the recently introduced EfficientNet models for the purpose of document classification and showed that parallelized multi-GPU training can improve classification accuracy, providing a new baseline for CNN-based models for document image classification. Recently, there has been an increased emphasis on multimodal classification techniques [30], [17], [18], in which images are preprocessed to extract the textual content of documents using standalone OCR software, and then visual, textual, and other layout information is combined to generate document representations and perform classification. Due to their recent success, vision transformers [31] have also gained some attention in document image classification [32], however, more work is needed before they can match the performance of the latest CNN-based models.

B. Model Robustness

There is rich literature on the subject of model robustness, which can be divided into two broad categories, namely adversarial robustness and corruption robustness. Adversarial robustness deals with the robustness of a model to an adversarial attack which refers to the introduction of a carefully crafted, imperceptible perturbation into a clean data input with the goal of confusing a machine learning classifier. Several types of adversarial attacks have been proposed in the past. For example, Su et al. (2017) [33] have shown how changing a single pixel in an input image can cause a deep learning classifier to fail. Similarly, Goodfellow et al. (2014) [13] have shown that very small, imperceptible additions to the image can affect the performance of black-box classifiers. To make the models more robust against these attacks, new defense mechanisms are constantly proposed in the literature [34].

While adversarial robustness focuses more on generating worst-case examples to confuse the machine learning models, corruption robustness is concerned with introducing minor corruptions or distortions into the data to determine how well the models can perform in real-world scenarios where these distortions are commonly found. Many studies have been conducted in this area showing the vulnerability of deep networks to these distortions. For example, Hosseini et al. (2017) showed that Google's Cloud Vision API can be easily fooled by adding impulse noise to input images. Geirhos et al. (2017) [35] and Dodge et al. (2017) [7] compared the performance of deep neural networks (DNNs) with that of humans and showed that they perform significantly worse than humans for various types of image distortions such as noise and blur, even when the networks were fine-tuned on those specific distortions.

To thoroughly investigate the robustness of modern deep learning models against common data distortions, Hendrycks et al. (2019a) [9] introduced robustness benchmark datasets for ImageNet [36]. They applied 15 different types of common visual perturbations with varying severity to ImageNet to create a new corrupted dataset, namely ImageNet-C, from which our work is directly inspired. They then evaluated the performance of modern deep learning networks on these datasets and showed that the classification error of these models increased significantly with only minor perturbations in the data. The two datasets have since been used as benchmarks for many robustness experiments and have allowed researchers to improve the robustness of existing neural networks, e.g., by introducing shape bias [37], extensive pretraining [38], or different types of data augmentation [39], [40].

III. RVL-CDIP-D and Tobacco3482-D Robustness Benchmarks

A. Base Datasets

In this section, we briefly describe the original datasets, namely, RVL-CDIP [14], and Tobacco3482, from which we have generated our robustness benchmarks. RVL-CDIP is one of the most well-known document image classification

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2022 doi:10.20944/preprints202202.0058.v1

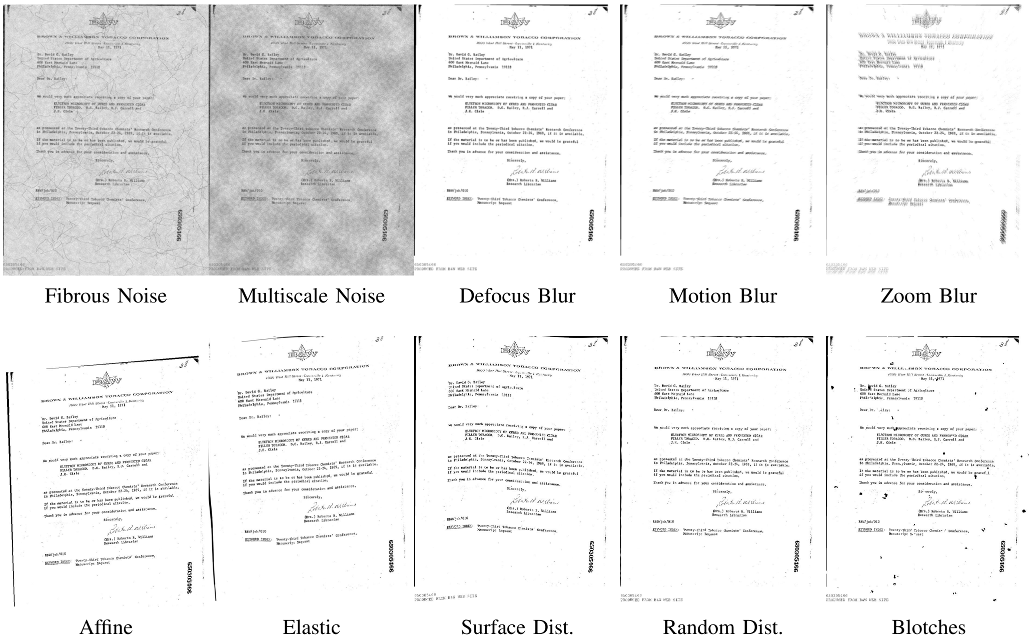

Fibrous Noise
Multiscale Noise
Defocus Blur
Motion Blur
Zoom Blur
Affine
Elastic
Surface Dist.
Random Dist.
Blotches

Fig. 1. A few of the different types of data distortions used in our robustness datasets are shown at severity level 3. Each type of distortion has five severity levels, giving a total of 105 different distortion types defined for each image.

datasets, which has been used as a benchmark in a number of studies [15], [19], [41], [17], [18]. The dataset contains a total of 400K labelled document images with 16 different document classes and is divided into the training, testing and validation sets of sizes 320K, 40K and 40K respectively. Tobacco-3482 is another popular dataset in the field of document image classification [28], [15] and contains a total of 3482 labelled document images. Since the dataset does not contain predefined splits, for our experiments, we split the dataset into training and testing sets with a ratio of 80/20.

B. Benchmarks Design

To create the robustness benchmarks, we followed the approach of [9] and defined a total of 21 different types of distortions drawn from the five main categories, namely noise, blur, geometric, digital, and document-specific distortions. A few of these distortions are illustrated in Fig. 1. Since in a real scenario distortions can occur in the data with different intensity, we also defined 5 different severity levels for each type of distortion, from least to worst. These distortions were then applied to the entire test set of the RVL-CDIP dataset to produce the RVL-CDIP-D dataset with a total of 4.2M images. Since there is no originally defined test set for the Tobacco3482 dataset, we simply applied the distortions to the entire dataset to create the Tobacco3482-D dataset. However, for the analysis, we used the portion of the dataset (73.5K images in total) that corresponded to the test split of the dataset, as defined in Section. III-A. The distortions were algorithmically defined using a combination of python packages: torchvision [1], ocrodeg [2], and opencv-python [3]. The dataset can be downloaded or recreated from our repository at https://github.com/saifullah3396/doc_robustness.

C. Distortions

In this section, we briefly describe the different types of distortions that we have used in this study.

Noise – *Gaussian noise* is often found in images taken in low light conditions. *Shot noise* is an electronic noise often found in images due to the discrete nature of light. *Fibrous noise* and *Multiscale noise* are document-specific noise and are used to represent the deterioration of the paper or its various textures that are common in the real world.

Blur – *Defocus blur*, *Motion blur*, and *Zoom blur* can all appear in images taken by a camera when it is out of focus, moving quickly, or zooming rapidly across the image. *Gaussian blur* is usually applied to images as a post-processing step when downsampling or upsampling images. *Binary blur* is another type of blur used in our study specifically to represent the spreading, erasure, and smudging of ink in documents. Similarly, *Noisy binary blur* is used to represent the spreading and bleed-through of ink.

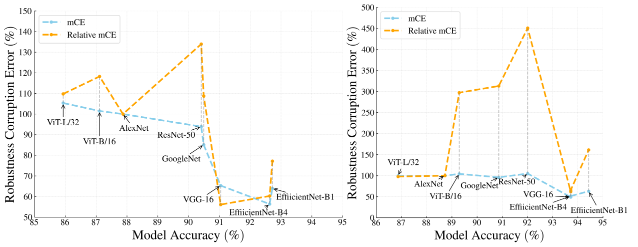
Digital – *Brightness* and *Contrast* may be high or low depending on the different lighting conditions. *Pixelation* may result from upsampling a low resolution image. *JPEG compression* is commonly used for lossy compression of digital images and results in certain compression artifacts.

Geometric – *Affine transformation* represents the translation, rotation and shear transformation of images with different intensities. This is important because documents in the real world are often slightly rotated, shifted, or warped. *Scale* is used to represent images taken by a camera from varying

[1] https://github.com/pytorch/vision
[2] https://github.com/NVlabs/ocrodeg
[3] https://github.com/opencv/opencv-python

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2022

doi:10.20944/preprints202202.0058.v1



(a) RVL-CDIP-D

(b) Tobacco3482-D

Fig. 2. A comparison of the Mean Corruption Error (mCE) and Relative mCE values of different networks with their accuracy.

distances. *Elastic transformation* stretches or contracts the images and can be used to represent the deformation and warping of the paper.

Document Specific – *Random Distortion* locally distorts the textual regions of the image and is used in our study to represent the spread and bleed-through of ink in document images. *Surface Distortion* is another type of distortion that represents the deformation or curling of the paper often found in document images. *Random blotches* randomly adds or removes blobs over the image to mimic the presence of ink blobs in documents or the erasure of ink from document paper. *Threshold* is a processing technique that we used in our study to mimic the degradation of textual information in documents over time.

D. Evaluation Metrics

To evaluate the performance of the different classifiers on our robustness benchmark datasets, we used the two evaluation metrics originally proposed by Hendrycks et al. (2019) [9], namely the mean corruption error (mCE) and the relative mean corruption error ($Rel.\ mCE$). Let $E^f_{s,c}$ be the error rate of a trained classifier f on data corrupted by distortion type c with severity s, then the mean corruption error mCE of classifier f is defined as the total classification error of f with respect to the baseline classifier on the distorted dataset and can be calculated as follows:

$$mCE^f = \frac{1}{n_c} \sum_{c=1}^{n_c} \left[(\sum_{s=1}^{n_{s,c}} E^f_{s,c}) / (\sum_{s=1}^{n_{s,c}} E^{f'}_{s,c}) \right] \quad (1)$$

Where f' represents the baseline classifier used to normalize the distortion errors, n_c denotes the total number of distortion types applied to the data, and $n_{s,c}$ denotes the number of severity levels defined for each distortion. The second evaluation metric, namely relative mCE, calculates the relative decline in the performance of a given classifier f when distortion is added and can be obtained by the following equation:

$$Rel.\ mCE^f = \frac{1}{n_c} \sum_{c=1}^{n_c} \left[(\sum_{s=1}^{n_{s,c}} E^f_{s,c} - E^f_{clean}) / (\sum_{s=1}^{n_{s,c}} E^{f'}_{s,c} - E^{f'}_{clean}) \right] \quad (2)$$

IV. Results

A. Models

In order to determine how well the latest document image classification models perform against common distortions, we evaluated several powerful deep neural networks available in the literature against the proposed robustness benchmarks. Afzal et al. (2017) [15] presented four deep models in their work, namely AlexNet, ResNet-50, VGG-16, and GoogleNet, which showed exceptional performance on both RVL-CDIP and Tobacco-3482 datasets. We selected all four models for evaluation in our work. More recently, Ferrando et al. (2020) [19] achieved new peak accuracy with different variants of the EfficientNet model [4] on both datasets. From their work, we selected two variants, EfficientNet-B1 and EfficientNet-B4, which were shown to perform best on the Tobacco3482 and RVL-CDIP datasets, respectively. In addition to these models, our study also examined the robustness of recently proposed vision transformers. Although it was shown by Siddiqui et al. [32] that it is difficult to train the existing vision transformers on the RVL-CDIP and Tobacco3482 datasets to achieve comparable performance to the latest CNN-based techniques, we thought it worthwhile to investigate their robustness to common distortions and therefore chose two of the commonly used variants, namely ViT-B/16 and ViT-L/32.

To reproduce the accuracy of the selected models, we followed the approach of [15] to first fine-tune the ImageNet pretrained models on the RVL-CDIP dataset and then further fine-tune them on the Tobacco3482 dataset. Since the RVL-CDIP and Tobacco3482 datasets are visually similar and contain some overlapping images, we removed the overlapping images from both datasets for transfer learning. To train the vision transformers, we fine-tuned the models using SGD (with a momentum of 0.9 and learning rates of 0.001-0.01), gradient clipping at norm 1, and a cosine learning rate schedule with linear warm-up. With this configuration of hyperparameters, we were able to slightly improve the performance of the vision transformers compared to that reported in [32]. After reproducing the accuracy of the models on the original datasets, we evaluated them on the proposed robustness benchmarks to compute the robustness metrics as defined in Section III-D. For all of our evaluation results, AlexNet was chosen as the baseline model.

B. Evaluation on Robustness Benchmarks

The results of our study are summarized in Table. I, which summarizes for each model the classification accuracy ($\text{Acc}_{\text{clean}}$) and error ($\text{Error}_{\text{clean}}$) on the original undistorted datasets, the mean error on each individual distortion type, and finally the mean corruption error (mCE) as defined in Section. III-D. Since AlexNet was chosen as the baseline network in our study, we also show the unnormalized errors of AlexNet separately so that the actual impact of each distortion type on classification performance can be compared. In addition, Fig. 2 shows a comparison of the total and relative corruption errors of the models with respect to their accuracy. A number

TABLE I

CLEAN ACCURACY, CLEAN ERROR, AND MCE, AND THE ERROR VALUES FOR DIFFERENT DISTORTION TYPES FOR EACH MODEL. ALEXNET (UNNORMALIZED) SHOWS THE ACTUAL MAGNITUDE OF THE ERRORS CAUSED BY THE DISTORTIONS. THE ERROR VALUES FOR TWO ADDITIONAL DISTORTION TYPES, I.E. GAUSSIAN BLUR AND NOISY BINARY BLUR ARE EXCLUDED.

							Noise								Blur							
Model	Acc_{clean}		E_{clean}		mCE		Gaussian		Shot		Fibrous		Multiscale		Defocus		Motion		Zoom		Binary	
	R	T	R	T	R	T	R	T	R	T	R	T	R	T	R	T	R	T	R	T	R	T
AlexNet (Unnormalized)	87.9	88.7	12.1	11.3	21.7	22.0	16	12	16	12	29	27	45	45	17	23	24	34	21	26	15	11
AlexNet [15]	87.9	88.7	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
VGG-16 [15]	91.0	93.7	9.0	6.3	65.32	49.0	59	55	59	58	54	45	55	43	57	31	44	23	53	35	75	59
GoogleNet [15]	90.5	90.9	9.51	9.1	85.10	96.0	88	127	88	127	131	167	122	135	88	98	84	90	95	93	91	74
ResNet-50 [15]	90.4	92.0	9.59	8.0	93.91	105.0	99	140	99	138	179	268	161	187	93	94	83	85	106	112	94	70
EfficientNet-B1 [19]	92.7	94.4	7.3	5.6	63.9	63.0	56	64	56	62	95	156	102	131	64	50	98	79	91	70	74	55
EfficientNet-B4 [19]	92.6	93.7	7.4	6.3	56.36	51.5	54	56	54	58	62	76	64	90	54	42	56	41	70	48	72	56
ViT-B/16 (Ours)	87.1	89.3	12.9	10.7	101.5	104.3	106	106	108	105	101	137	90	95	102	107	79	81	103	97	126	100
ViT-L/32 (Ours)	85.9	86.9	14.1	13.1	105.4	99.6	105	117	105	110	116	114	126	118	96	85	71	63	84	71	123	105
					Mean	=	83	96	84	95	105	133	103	112	82	76	77	70	88	78	94	77

	Digital								Geometric						Documents Specific							
Model	Brightness		Contrast		Pixelate		JPEG		Affine		Scale		Elastic		Surf Dist.		Rand Dist.		Blotches		Threshold	
	R	T	R	T	R	T	R	T	R	T	R	T	R	T	R	T	R	T	R	T	R	T
AlexNet (Unnormalized)	19	18	37	45	12	11	12	12	25	23	25	22	16	17	12	11	12	12	23	21	15	11
AlexNet [15]	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
VGG-16 [15]	61	42	59	30	73	56	74	54	84	72	80	64	69	50	74	61	60	55	77	44	74	63
GoogleNet [15]	65	77	73	78	80	98	83	103	74	68	76	78	76	93	80	99	80	93	60	55	87	90
ResNet-50 [15]	63	75	104	107	81	94	87	103	76	69	77	72	81	85	80	91	81	93	67	57	90	89
EfficientNet-B1 [19]	44	38	37	38	61	59	63	58	51	49	58	51	56	49	60	59	61	53	39	38	62	56
EfficientNet-B4 [19]	45	38	32	30	63	56	63	60	50	44	58	46	56	44	61	58	62	54	40	33	61	61
ViT-B/16 (Ours)	88	86	63	64	108	118	109	114	104	102	103	99	115	120	115	119	109	178	85	61	117	105
ViT-L/32 (Ours)	99	88	73	56	115	114	117	123	111	101	108	96	124	116	129	122	117	120	92	68	117	121
Mean =	71	68	68	63	85	87	87	90	81	76	83	76	84	82	87	89	86	93	68	57	89	86

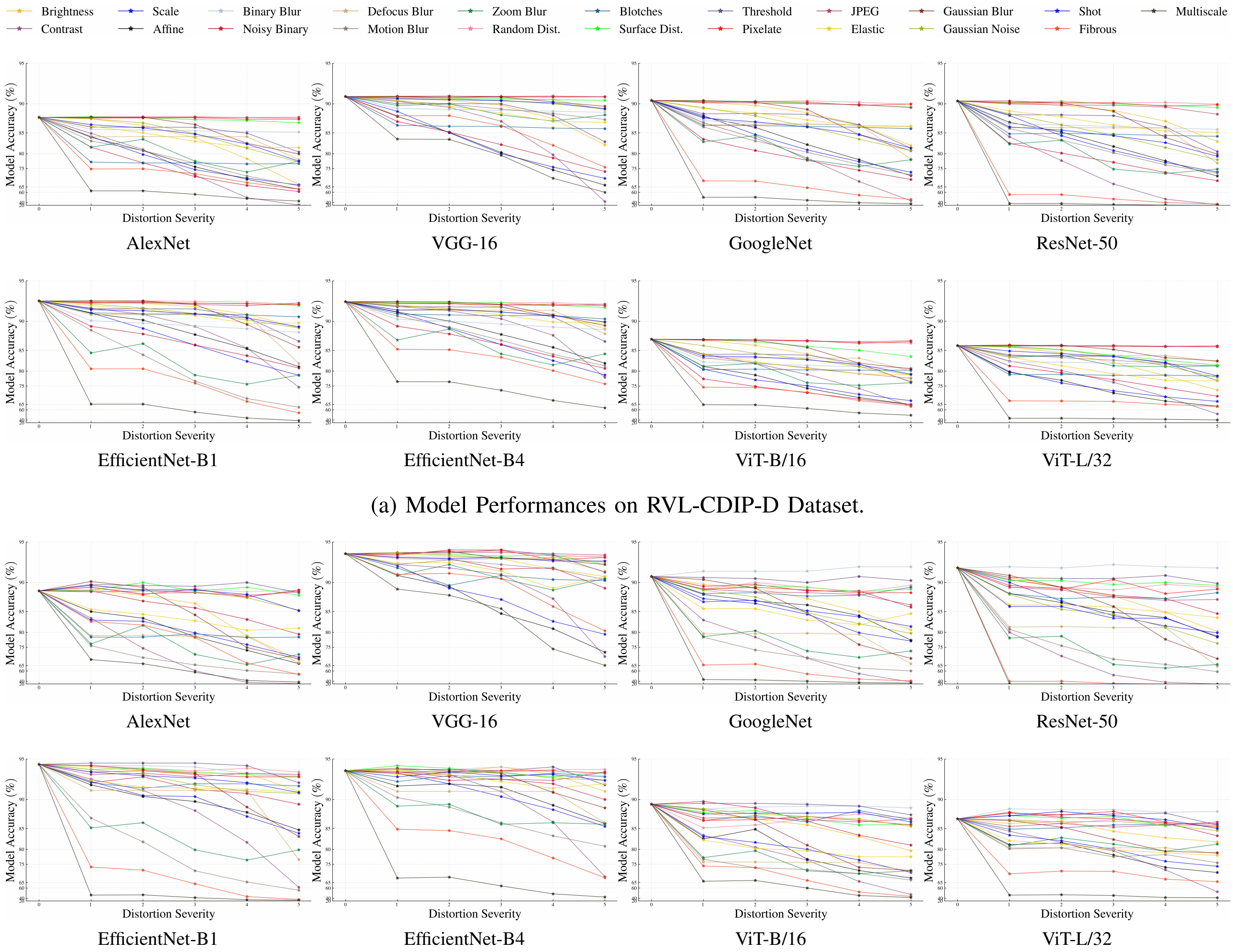
R: RVL-CDIP, T: Tobacco3482

of conclusions can be drawn from these results. For instance, it can be seen that with increasing accuracy of the models, their mean robustness to common distortions has generally increased. However, the relative increase in distortion errors is much higher for some models. For example, ResNet-50 and GoogleNet had comparable accuracy to the VGG-16 model, but performed comparatively much worse in terms of robustness. It can also be seen that the relative performance degradation of ResNet-50 and GoogleNet was exceptionally high on both datasets, reaching 300-450% that of AlexNet. On the other hand, the EfficientNet variants both performed exceptionally well in terms of accuracy and robustness. However, the relative robustness of these classifiers was still lower than that of VGG-16 on both datasets. Finally, we can see that while vision transformers were able to achieve comparable performance to AlexNet, they still lag slightly behind AlexNet in terms of robustness. Nevertheless, ViT-L/32 model was still seen to perform better than ResNet-50 and GoogleNet in terms of relative robustness.

We can also make some important observations about which types of corruption most affect the performance of the classifiers. First, we can see from the tables that multiscale noise and contrast distortions had the greatest impact on the performance of both CNNs and ViTs, with the mean unnormalized error for AlexNet reaching up to 45%. However, the VGG-16 and EfficientNet variants were still able to perform relatively well with these types of distortions. Of the different types of blur, motion blur seemed to have the greatest impact, followed by defocus blur and zoom blur. Affine and scale distortions, as well as document-specific random blotches also appeared to have a significant impact on the overall performance of the classifiers. In contrast, digital distortions had less of an overall impact, with the exception of contrast. Overall, on the RVL-CDIP-D dataset, EfficientNet-B4 showed the highest robustness to all types of distortions, followed by EfficientNet-B1 and VGG-16. On Tobacco3482-D dataset, on the other hand, VGG-16 performed best among all models, followed by EfficientNet-B4 and EfficientNet-B1.

We also illustrate the effect of increasing severity of each distortion type on the performance of the models in Fig. 3, from which some important observations can be drawn. For example, it can be seen that the CNN-based models were consistently robust to some specific distortions such as surface distortion, random distortion, JPEG compression, and pixelation. Other distortions such as binary blur, defocus blur, Gaussian noise, and shot noise, etc. caused a moderate decline in accuracy that increased with severity. Similar to our previous observation, motion blur, contrast, fibrous noise, and multi-scale noise had a strong effect on the performance of the models, and the effect increased with severity. In addition to analyzing the effects of individual distortions, we can also draw meaningful conclusions by comparing different models. For example, when we compare the results of VGG-16 and EfficientNet-B4 on the RVL-CDIP-D dataset, we find

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2022 doi:10.20944/preprints202202.0058.v1



(a) Model Performances on RVL-CDIP-D Dataset.

(a) Model Performances on Tobacco3482-D Dataset.

Fig. 3. Shows the effect of increasing severity on model performance for each distortion type.

that the performance of VGG-16 for severity > 3 decreased significantly for some distortions, while EfficientNet-B4 still showed a comparatively small decrease in accuracy. This is consistent with our findings on the mean corruption error of VGG-16, which is slightly higher for RVL-CDIP-D than for EfficientNet-B4. Overall, we can conclude that VGG-16 and EfficientNet-B4 are indeed quite robust to these distortions (except for a few), at least up to severity level 3. Another interesting observation is that for ViT-L/32, the decrease in accuracy is much smaller for many distortions compared to the GoogleNet and ResNet-50 models, which supports our previous findings about the model from Fig. 2.

V. Conclusion

In this paper, we presented two novel benchmark datasets for evaluating the robustness of document image classifiers to common corruptions. In addition, we evaluated the current state-of-the-art deep learning classifiers on these datasets to analyze how well they perform on corrupted datasets. Through the analysis, we found that while some of the latest techniques are quite robust to corruption, a few others can perform even worse than AlexNet. This result is important because it shows that even if two deep networks show similar performance in terms of accuracy, they may differ significantly in terms of robustness and therefore accuracy alone should not be sufficient to evaluate the overall strength of deep classifiers. We also present comprehensive results on the impact of each corruption type on classifier performance, which may help future research to improve the robustness of these classifiers to specific corruption types. A plausible future work can be to assess the performance of the latest robustness improvement techniques on document image classifiers using our proposed benchmarks. Another future direction might be to use these datasets to evaluate the robustness of existing multi-modal document classification techniques to common real-world distortions.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[4] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 6105–6114.

[5] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "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. Association for Computational Linguistics (ACL), 2019, pp. 4171–4186.

[6] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 6645–6649.

[7] S. Dodge and L. Karam, "A study and comparison of human and deep learning recognition performance under visual distortions," *2017 26th Int. Conf. Comput. Commun. Networks, ICCCN 2017*, 2017.

[8] H. Hosseini, B. Xiao, and R. Poovendran, "Google's cloud vision API is not robust to noise," *Proc. - 16th IEEE Int. Conf. Mach. Learn. Appl. ICMLA 2017*, vol. 2017-December, pp. 101–105, 2017.

[9] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," *7th Int. Conf. Learn. Represent. ICLR 2019*, pp. 1–16, 2019.

[10] Z. Lipton, Y.-X. Wang, and A. Smola, "Detecting and correcting for label shift with black box predictors," 02 2018.

[11] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," 10 2016.

[12] N. Carlini and D. Wagner, "Adversarial examples are not easily detected: Bypassing ten detection methods," *AISec 2017 - Proc. 10th ACM Work. Artif. Intell. Secur. co-located with CCS 2017*, pp. 3–14, 2017.

[13] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.* International Conference on Learning Representations, ICLR, dec 2015.

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

[15] M. Z. Afzal, A. Kolsch, S. Ahmed, and M. Liwicki, "Cutting the Error by Half: Investigation of Very Deep CNN and Advanced Training Strategies for Document Image Classification," *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, vol. 1, pp. 883–888, 2017.

[16] M. Z. Afzal, S. Capobianco, M. I. Malik, S. Marinai, T. M. Breuel, A. Dengel, and M. Liwicki, "Deepdocclassifier: Document classification with deep Convolutional Neural Network," *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, vol. 2015-Novem, pp. 1111–1115, 2015.

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

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

[19] J. Ferrando, J. L. Domínguez, J. Torres, R. García, D. García, D. Garrido, J. Cortada, and M. Valero, "Improving accuracy and speeding up document image classification through parallel systems," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12138 LNCS, pp. 387–400, 2020.

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

[21] D. Shin, Christian and Doermann, "Document Image Retrieval Based on Layout Structural Similarity." *Proc. 2006 Int. Conf. Image Process. Comput. Vision, Pattern Recognit.*, vol. 2, pp. 606–612, 2016.

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

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

[24] K. Collins-Thompson and R. Nickolov, "A Clustering-Based Algorithm for Automatic Document Separation," *Proc. SIGIR 2002 Work. Inf. Retr. OCR From Convert. Content to Grasping Mean.*, no. September 2002, 2002.

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

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

[27] N. Chen and D. Blostein, "A survey of document image classification: Problem statement, classifier architecture and performance evaluation," *Int. J. Doc. Anal. Recognit.*, vol. 10, no. 1, pp. 1–16, 2007.

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

[29] A. Das, S. Roy, U. Bhattacharya, and S. K. Parui, "Document Image Classification with Intra-Domain Transfer Learning and Stacked Generalization of Deep Convolutional Neural Networks," *Proc. - Int. Conf. Pattern Recognit.*, vol. 2018-Augus, pp. 3180–3185, 2018.

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

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

[32] S. A. Siddiqui, A. Dengel, and S. Ahmed, "Analyzing the potential of zero-shot recognition for document image classification," in *Document Analysis and Recognition – ICDAR 2021*, J. Lladós, D. Lopresti, and S. Uchida, Eds. Cham: Springer International Publishing, 2021, pp. 293–304.

[33] J. Su, D. V. Vargas, and K. Sakurai, "One Pixel Attack for Fooling Deep Neural Networks," *IEEE Trans. Evol. Comput.*, vol. 23, no. 5, pp. 828–841, 2019.

[34] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial Examples: Attacks and Defenses for Deep Learning," *IEEE Trans. neural networks Learn. Syst.*, vol. 30, no. 9, pp. 2805–2824, 2019.

[35] R. Geirhos, D. H. J. Janssen, H. H. Schütt, J. Rauber, M. Bethge, and F. A. Wichmann, "Comparing deep neural networks against humans: object recognition when the signal gets weaker," 2017.

[36] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[37] R. Geirhos, C. Michaelis, F. A. Wichmann, P. Rubisch, M. Bethge, and W. Brendel, "Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," *7th Int. Conf. Learn. Represent. ICLR 2019*, no. c, pp. 1–20, 2019.

[38] D. Hendrycks, K. Lee, and M. Mazeika, "Using pre-training can improve model robustness and uncertainty," *36th Int. Conf. Mach. Learn. ICML 2019*, vol. 2019-June, no. 2018, pp. 4815–4826, 2019.

[39] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty," pp. 1–15, 2019.

[40] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, D. Song, J. Steinhardt, and J. Gilmer, "The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization," 2020.

[41] C. Tensmeyer and T. Martinez, "Analysis of Convolutional Neural Networks for Document Image Classification," *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, vol. 1, pp. 388–393, 2017.