BusiNet - a Light and Fast Text Detection Network for Business Documents

Oshri Naparstek
oshri.naparstek@ibm.com
IBM Research - Haifa
Israel

Ophir Azulai
ophir@il.ibm.com
IBM Research - Haifa
Israel

Daniel Rotman
danieln@il.ibm.com
IBM Research - Haifa
Israel

Yevgeny Burshtein
bursh@il.ibm.com
IBM Research - Haifa
Israel

Peter Staar
taa@zurich.ibm.com
IBM Research - Zurich
Switzerland

Udi Barzelay
udib@il.ibm.com
IBM Research - Haifa
Israel

ABSTRACT

For digitizing or indexing physical documents, Optical Character Recognition (OCR), the process of extracting textual information from scanned documents, is a vital technology. When a document is visually damaged or contains non-textual elements, existing technologies can yield poor results, as erroneous detection results can greatly affect the quality of OCR.

In this paper we present a detection network dubbed BusiNet aimed at OCR of business documents. Business documents often include sensitive information and as such they cannot be uploaded to a cloud service for OCR. BusiNet was designed to be fast and light so it could run locally preventing privacy issues. Furthermore, BusiNet is built to handle scanned document corruption and noise using a specialized synthetic dataset. The model is made robust to unseen noise by employing adversarial training strategies. We perform an evaluation on publicly available datasets demonstrating the usefulness and broad applicability of our model.

CCS CONCEPTS

• **Information systems → Document structure**; • **Applied computing → Document analysis**; *Optical character recognition*; • **Computing methodologies → Object detection**.

KEYWORDS

Text Detection, Document Analysis,adversarial training, synthetic data, Unet, Segmentation

1 INTRODUCTION

Documents have always been, and continue to be, a significant data source for any business or corporation. For physical documents, the ability to scan and digitize them is crucial in order to extract their information and represent them in a way that allows for further analysis. To this day, the capability to automatically ingest and read scanned IDs, invoices, itineraries, statements, and spreadsheets remains an enormously important technology.

In many instances these documents are scanned or digitally acquired under non-ideal conditions. These include incorrect scanner settings, insufficient resolution, bad lighting, loss of focus, unaligned pages, and added artifacts from badly printed documents. Thus a high quality version of the document is unattainable while manual annotation or error correction is costly.

Besides this, a major limitation for many corporations when analyzing documents is the need to protect sensitive or proprietary information contained in the documents. Using popular cloud services is often undesirable or impossible, as exporting documents to external technologies to be analysed can often seem like a compromising and daunting endeavour. On the other hand, many businesses lack the infrastructure to run popular document analysis technologies at the necessary quantity and speed.

To perform the digitization of documents, Optical Character Recognition (OCR) is utilized. OCR is composed of a detection stage where the various words in the document are localized, and a recognition stage to identify the comprising characters.

In this work we present BusiNet - a lightweight detection network for document OCR. Our minimalist U-Net architecture promotes fast and accurate detection trained specifically for the document domain. Extending the CRAFT [1] methodology of multiple output channels, and utilizing adversarial training, we can create a powerful and robust detector.

Powering the trained detection network is our composition of new data synthesis components. Extending the work from [9], we generate specific elements that characterise documents and their expected noise and artifacts. This includes elements such as separators, lines, special characters, and punctuation. For document-wide alterations we apply a host of spatial distortions, augmentations, and backgrounds. Our main contributions are as follows: We use very light and fast backbone so the model could operate locally. We separate the output channels into letters, spaces between letters and special characters to better handle separation between words. We use specialized synthetic dataset to capture common document corruption. To make it more robust against unknown noise, we employ adversarial training strategy.

2 PREVIOUS WORK

2.1 Detection

U-Net-based segmentation maps are very common for detecting multiple objects in images [8]. Essentially, the architecture relies on convolutional layers, which reduce the spatial element to a bottleneck, and then up convolutions, which restore the semantic information to its original spatial dimensions. U-Net architecture and variations can be applied to a variety of fields, including image segmentation (which is widely used in medical imaging) [5, 6, 15], in addition to other tasks such as saliency detection [3], or as GAN discriminators [10].

Many powerful text detectors are constructed with architectures to promote STR (Scene Text Recognition), i.e., isolating text in natural images such as billboards or street signs. EAST [14] features a contracting and expanding network similar to the U-Net, and performs regression on the quadrilaterals based on the feature which is also used to generate the score map. CRAFT [1] also use a U-Net architecture as a backbone. In CRAFT, the output is split between character detection and spaces between characters detection. PAN++ [12] which is an improvement over PAN [13] adopts a Feature Pyramid Enhancement Module which is similar to a concatenation between two U-Nets. Detection networks for OCR are less popular in recent academic studies, and even more so when attempting to find recent methods for comparison.

3 METHOD

Our goal is to provide a text detection model which is best suited for business documents. This means that the model should be lightweight so it can run on-premise and also be accurate on low quality scans and faxes. In this section we present the BusiNet model architecture designed to achieve these goals. BusiNet is built around a lightweight U-Net [8] architecture with custom output channels and a custom-made specialized dataset. This makes the model suitable for the task of OCR of business documents done on-premise . It is then trained using an adversarial training strategy to make it more robust against unseen noise patterns and to attain accuracy on a wide variety of document types. A figure of the full detection pipeline of BusiNet is shown in Figure 1.

3.1 Data Synthesis

We train our model on synthetic data only. In [9] the authors describe a carefully designed synthetic dataset which increases the detection quality significantly. In this work we build on the results in [9] by using the proposed generator and improving it by adding advanced capabilities. This makes the generator more flexible and more compatible for adjustments for even more challenging documents. The fact that we use synthesized data enables us to tweak it and to add data that will help the model to preform better on a specific domain or type of noise. Business documents often contain graphical lines and special fonts. We added these types of text to our generator. We also add noise patterns which are characteristic to scanned business documents such as small lines and dots. An example of some of the new capabilities added to the generator are shown in Figure 2.

3.2 Lightweight backbone

U-Net [8] models have been shown to have good performance on semantic segmentation tasks. Hence, we adopt a U-Net architecture as our backbone. We use four layers of convolution, and then up-convolutions are performed with the first layer having 16 channels and the rest of the layers having 32 channels. In the up-convolution process, skip-connections are employed by concatenating the output feature of the up-convolution with the feature of the regular convolution at the same level.

3.3 Dealing with special characters

It is quite common in business documents for special characters to be used as separators. Depending on the context, these characters should be detected in some cases and ignored in others.

To address this, we leverage the architecture from CRAFT [1]. When dealing with detecting words, they suggested to separate the output into two channels, a channel that detects characters and a channel that detects spaces between adjacent characters. The reasoning behind this is that a character is better defined than a word, and this definition differentiates much more clearly between a word and a line. For this reason separating characters and spaces between characters yields better detection accuracy.

Following this methodology, we also separate the detection of characters from the detection of spaces between characters. However, we extend this method to mitigate the special character problem. As noted before, special characters are characters that need special treatment because of their dual purpose. To correctly deal with special characters, we add a third output channel in addition to the existing two dedicated to detecting special characters. After the detection is made, the special characters are combined with the other channels only if they are close to regular characters. This way we prevent special characters which are used as separators to be detected as text. This approach could be further extended to other types of texts or characters that require special attention. We leave this extension for a future work . An example for the effectiveness of the special character channel is shown in Figure 3.

3.4 Adversarial training

Our goal is for the text detection model to be as accurate as possible in the broadest range of scenarios as possible. In other words, the detection model should be able to detect corruption and noise that it has not been trained on. To make our model more robust against unknown image noise distributions and corruptions we trained our model using an adversarial training strategy in addition to the noise added during data synthesis.

Adversarial training is a method where during training, in addition to the original samples, the model is also trained with perturbed images. These samples are designed such that they pose the hardest task for the current state of the model.

As was shown in [11], neural networks are sensitive to adversarial attacks. Adversarial attacks are a small, carefully engineered perturbation to the image that causes the model to incorrectly classify it. Originally, adversarial training was used to make a model more robust to these adversarial attacks. A simple way to make the model more robust to adversarial attacks is to train it using samples that were perturbed in an adversarial approach.

We use Projected Gradient Descent (PGD) with the L_2 norm to create the adversarial examples. On a given input, PGD attempts to find the perturbation that maximises the model's loss while the perturbation size, denoted by ϵ, is kept below a certain amount. This is done in an iterative manner by taking a gradient step towards the direction of the greatest loss. If the perturbation is too large

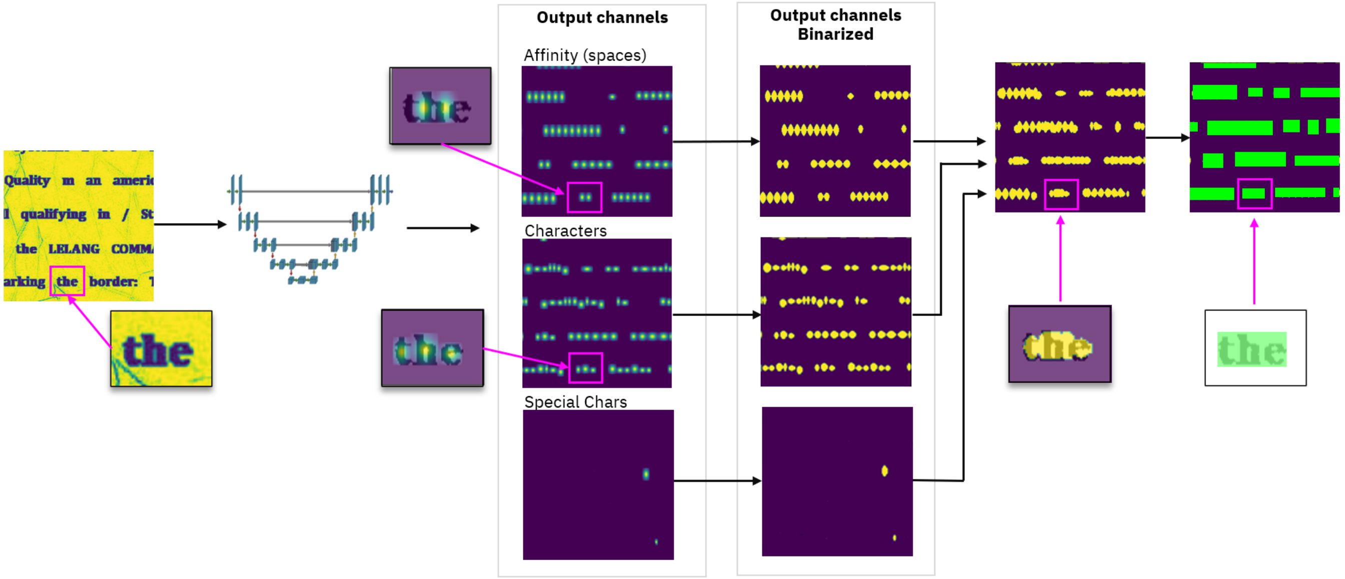


Figure 1: BusiNet pipeline: First, the text image goes through a U-Net with three output channels, a channel for character segmentation, a channel for spaces segmentation, and a channel for special character segmentation. A threshold is applied for each channel separately. Then the channels are combined while taking into account the special characters, and the word-level bounding boxes are constructed.

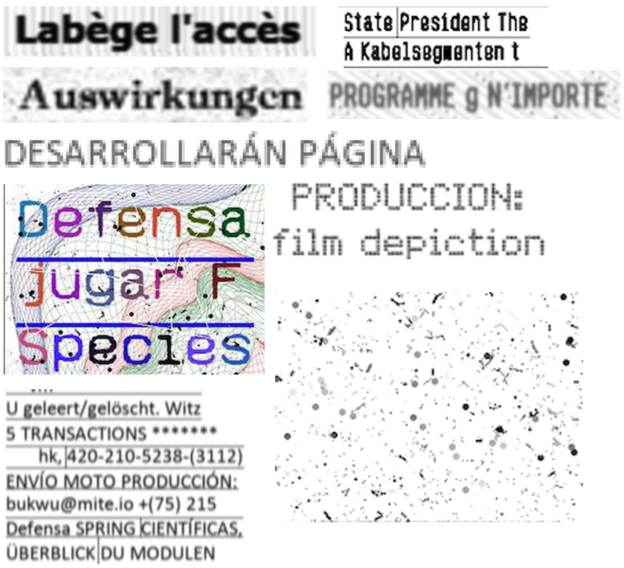


Figure 2: New styles added to the generator

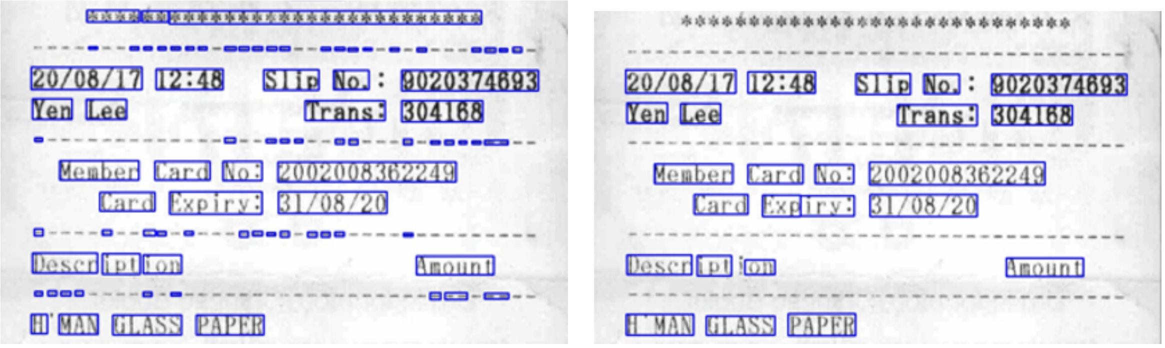
then it is projected back into the ball of allowed perturbations. In this work, the use of adversarial examples during training is done for a different reason than in other works. Adversarial examples force the network to train on samples which are harder for the current model weights. If a model is robust to adversarial attacks with energy less then some threshold ϵ it implies that it is also robust to any perturbation of the image which is smaller than ϵ. The reason for this is because the adversarial example is the worst case scenario for the model weights.

Also, as shown in [2], adversarial training tends to align the model features with the human perception making the model more interpretable. In Figure 4 we show an example to the influence of adversarial attack on a model that was not trained adversarially. The mild perturbation to the image results in a complete failure of the regular model while the adversarially trained model still preforms well.

4 EVALUATION

We evaluate our detection model on the SROIE [4] and FUNSD [7] datasets. We use the test sets with annotated text bounding boxes and transcribed words. We aim to compare between the methods in a setting that resembles real conditions under which the system will operate. For that reason we perform the comparison on one core of Intel(R) Core(TM) i7-5930K CPU @ 3.50GHz with 2GB of RAM which in our opinion is a reasonable representation of an on-premise setup. In Table 1 we present the results of our evaluation. We measure detection using the F1-score of correctly detected bounding boxes using IOU 0.5.

As shown in Table 1, BusiNet achieves a much higher F1 score compared to PAN++ and EAST. The performance is comparable to CRAFT while being almost 3 times faster than CRAFT. The difference becomes more apparent when comparing the OCR results between the methods. In Table 2 we compare the recognition



(a) (b)

Figure 3: Example to the effect of the special character channel on the detection accuracy. (a) An output of a detection model trained without the special character channel. (b) An output of a detection model trained with the special character channel.

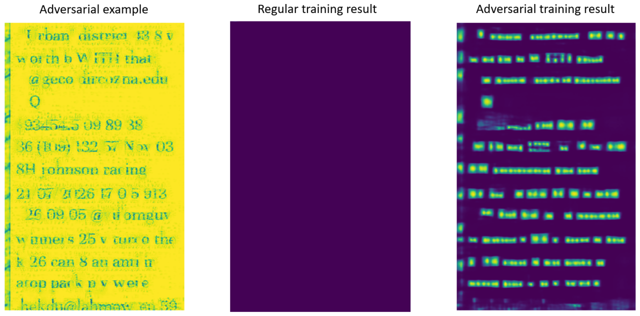


Figure 4: Comparison between detection results between regular training and adversarial training. The network that was trained in a standard approach completely fails while the adversarially trained network is almost unaffected by the perturbation.

accuracy of the Tesseract OCR engine on the detected text. The difference is even more pronounced where BusiNet is still comparable to CRAFT but greatly outperforms EAST, and PAN++.

These results demonstrate the usefulness of our method. It achieves accuracy comparable to much larger architectures yet runs much faster which means it can give fast and accurate results while running on-premise.

Table 1: Results . Detection measured by F1-score

Method	FUNSD F1	SROIE F1	Time (sec.)
PAN++ [12]	90.5	89.1	4.8
EAST [14]	82	79	2.8
CRAFT [1]	**97.6**	95.3	13.3
BusiNet - Ours	96.4	**95.9**	4.8

5 CONCLUSIONS

In this work, we presented a lightweight text detection model suited for noisy business documents. We try to address the main requirements of business documents costumers. The light backbone makes the inference time faster than other methods and allows for the model to run on-premise. The synthetic data and special characters channel improve the model robustness to common characteristics of business documents. Adversarial training ensures that the model will be as robust as possible to all types of noise known or unknown. We compare our results to popular and established text detection models and show that our model is more accurate than the other methods while being much lighter and faster.

REFERENCES

[1] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. 2019. Character region awareness for text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9365–9374.

[2] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. 2019. Adversarial robustness as a prior for learned representations. *arXiv preprint arXiv:1906.00945* (2019).

[3] Le Han, Xuelong Li, and Yongsheng Dong. 2019. Convolutional edge constraint-based U-net for salient object detection. *IEEE Access* 7 (2019), 48890–48900.

[4] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. 2019. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 1516–1520.

[5] Nabil Ibtehaz and M Sohel Rahman. 2020. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural Networks* 121 (2020), 74–87.

[6] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, et al. 2018. nnu-net: Self-adapting framework for u-net-based medical image

Table 2: Results . Recognition measured by average case-insensitive Edit Score (ES). 'Recognition' indicates using the method's detection for word-level recognition.

Method	FUNSD ED (Normalized)	SROIE ED (Normalized)
PAN++ [12]	69.5	43.8
EAST [14]	50.6	22.9
CRAFT [1]	**81.6**	52
BusiNet - Ours	79.2	**56.9**

segmentation. *arXiv preprint arXiv:1809.10486* (2018).

[7] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, Vol. 2. IEEE, 1–6.

[8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

[9] Daniel Rotman, Ophir Azulai, Inbar Shapira, Yevgeny Burshtein, and Udi Barzelay. 2022. Detection Masking for Improved OCR on Noisy Documents. *arXiv preprint arXiv:2205.08257* (2022).

[10] Edgar Schonfeld, Bernt Schiele, and Anna Khoreva. 2020. A u-net based discriminator for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8207–8216.

[11] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

[12] Wenhai Wang, Enze Xie, Xiang Li, Xuebo Liu, Ding Liang, Yang Zhibo, Tong Lu, and Chunhua Shen. 2021. PAN++: towards efficient and accurate End-to-End spotting of arbitrarily-shaped text. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[13] Wenhai Wang, Enze Xie, Xiaoge Song, Yuhang Zang, Wenjia Wang, Tong Lu, Gang Yu, and Chunhua Shen. 2019. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8440–8449.

[14] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. 2017. East: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 5551–5560.

[15] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. 2018. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 3–11.