Evaluation of nnU-Net method for Document Image Binarization

Academic and Research Laboratory in Project Computer Vision

submitted
by

Adrián Bécares Cantón

born 18.01.1999 in Vitoria-Gasteiz

Written at

Lehrstuhl für Mustererkennung (Informatik 5)
Department Informatik
Friedrich-Alexander-Universität Erlangen-Nürnberg.

Advisor: Dr.-Ing. Vincent Christlein

Started: 01.11.2022

Finished: 31.03.2023

Ich versichere, dass ich die Arbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat und von dieser als Teil einer Prüfungsleistung angenommen wurde. Alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.

Die Richtlinien des Lehrstuhls für Studien- und Diplomarbeiten habe ich gelesen und anerkannt, insbesondere die Regelung des Nutzungsrechts.

Erlangen, den 30. März 2023

Übersicht Die Binarisierung von Dokumentenbildern ist ein immer wiederkehrendes Problem bei wichtigen Aufgaben wie der optischen Zeichenerkennung. Die Einführung von Deep Learning für Segmentierungs- und Binarisierungszwecke hat die Leistung bei diesen Problemen erheblich gesteigert. Dieses Projekt zielt darauf ab, die Arbeit von Sukest et al. [Suk+22] zu erweitern, indem die neuartige Segmentierungspipeline nnU-Net [Ise+21] analysiert wird. Die Experimente wurden nach den gleichen Richtlinien wie in der oben genannten Arbeit durchgeführt, einschließlich der Verwendung der gleichen DIBCO-Challenge-Datensätze der Jahre 2009-2019 und der gleichen Bewertungsmetriken. Die Ergebnisse zeigen, dass die Leistung von nnU-Net im Vergleich zu den Gewinnern des DIBCO-Wettbewerbs und zu den besten in der vorherigen Arbeit analysierten Methoden schlecht ist, aber die Leistung der DRD-Bewertungsmetrik liegt über dem Durchschnitt.

Abstract Document Image Binarization is a recurring problem for essential tasks like Optical Character Recognition. The introduction of Deep Learning for segmentation and binarization purposes has greatly increased the performance in these problems. This project aims to extend the work of Sukest et al. [Suk+22] by analizing the novel segmentation pipeline nnU-Net [Ise+21]. The experimentation was performed following the same guidelines specified in the aforementioned work, including the usage of the same DIBCO challenge datasets of years 2009-2019 and the same evaluation metrics. Results show that the performance of nnU-Net is poor when compared to the winners of the DIBCO challenge and to the best methods analyzed in the previous work, but the performance of the DRD evaluation metric is above the average.

Contents

Chapter 1

Introduction

Document Image Binarization (DBI) has been a recurring problem since the beginning of Computer Vision and Pattern Recognition development [Suk+22; Ten+20] In this process, an RGB image is converted into a grayscale, binary image, where the pixels corresponding to text are assigned as black (foreground), and the pixels corresponding to the paper as white (background). This task is particularly useful for preprocessing steps when dealing with documents, for example, in Optical Character Recognition [Isl+17; Li+21] or writing and layout analysis [Chr+17b; Chr+17a].

Traditional binarization methods rely on statistical features to select the appropriate threshold that separate the desired text from the background, as well as eliminating noise or artifacts (e. g., stains) that may be within the document. Classic thresholding algorithms, such as Otsu's [Ots79], derives a global threshold from the histogram of a grayscale image minimizing the intra-class variance (the text) and maximizing the inter-class (text and background) variance. Although this method is widely used due to its effectiveness and low computational requirements, it is necessary to preprocess the image and convert it reliably into a grayscale image.

Modern technologies such as Deep Learning (DL), and specifically Convolutional Neural Networks (CNNs), have overcome this problem by training directly into the underlying features without the explicit requirement of selecting a global or multiple local thresholds. This ultimately acts as an end-to-end system capable of preprocess and binarize RGB images without more in-depth human intervention [Ron+15; Ise+21].

Other implementation of DL capabilities fall into the segmentation processes. Segmentation problems aim to separate and label individual pixels to certain objects or structures, this is, semantic labelling [Du+20; Min+22; Ron+15; Ise+21]. This problem usually tackles

a wide range of applications, such as medical image segmentation (e. g., localizing and separating tumor cells from healthy cells), autonomous vehicles, enhanced or augmented reality, etc. In this field, U-Net architectures are a common choice due to their excellent performance with very few training examples, which is a usual problem in many medical datasets. In addition, recent methods such as nnU-Net [Ise+21], aim to analyze the used dataset in order to dynamically adapt its architecture in an end-to-end pipeline without further intervention.

Although there exist several methods for document image binarization, few research has been performed to detect which of these techniques performs best when trained in such applications. This was overcomed by the work of Sukesh et al. [Suk+22], which trained several commonly-used binarization methods on the Document Image Binarization Contests (DIBCO) [Pra+09; Pra+10; Pra+11; Pra+12; Pra+13; Pra+14; Pra+16; Pra+17; Pra+18; Pra+19] datasets from 2009 to 2019 and evaluated them using 4 metrics. The aim of this project is to extend the research performed by Sukesh et al. [Suk+22] by training the novel nnU-Net(v1) architecture with the same data and performing the same evaluation steps used in the aforementioned work.

The following report explains the functioning of the analyzed nnU-Net architecture, as well as the implementation, performed experimentation and result analysis. These last show that the usage of this novel architecture does not perform better than the evaluated methods in Sukesh et al. [Suk+22]. Nevertheless, some future extensions, like increasing the time-span for training and the implementation and re-training of nnU-Net(v2) may increase the performance of the architecture and thus introduce segmentation-oriented methods into the document binarization context.

Chapter 2

Evaluated Architecture

2.1 Fully Convolutional Networks and U-Net Architecture

Traditional usage of CNNs has involved the processing of an image in subsequent convolutional block steps and a final fully-connected layer that actuates as a classifier into a single label [Ron+15]. However, segmentation problems usually require of multiple pixel labeling (localization), thus requiring for the introduction of "fully convolutional" [Lon+15] idea, in which the output classification is a mapping combination dependent on the number of required labels. Here, segmentation can be seen as a generalization of the binarization problem, where the number of classes that the image must be segmented to is 2.

U-Net [Ron+15] extends this approach by creating an architecture consisting on two major steps, contracting and expansive paths. Figure 2.1 shows the typical architecture of these networks. The left side or contractive path involves the standard repetition of convolutions to achieve a latent representation of the image. The expansive path up-samples this obtained feature vector with the so-called "up-convolutions" concatenated with the corresponding feature vector of the contractive path to avoid the loss of border pixels. Finally, a 1x1 convolution is performed to obtain the desired number of segmentations.

2.2 nnU-Net: self configuring method and pipeline

Although the U-Net architecture obtains great performance in different biomedical segmentation applications, it still requires of expertise and experience for the correct configuration

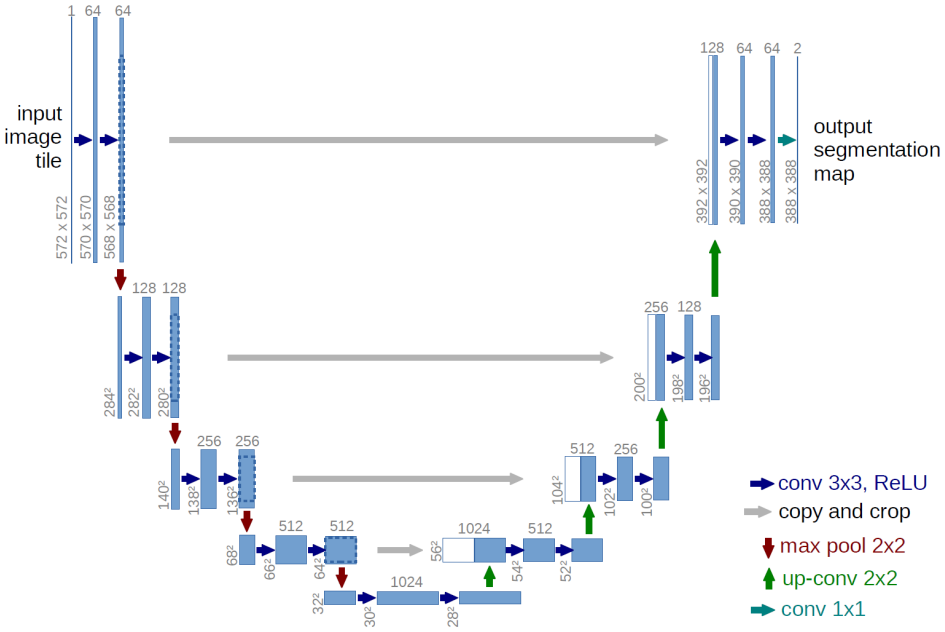


Figure 2.1: Architecture of the U-Net proposed by Ronnenberger et al. [Ron+15]. This example shows the resulting 2 segmentation maps that a binarization procedure will produce.

of the hyperparameters. These configurations vary extremely from one dataset to another [Ise+21]. Thus, the introduction of a self-configuring architecture is necessary for the automatic selection of parameters in order to reduce the process time consumption.

The analyzed nnU-Net architecture systematizes the configuration procedure by collecting design decisions that do not require changes between different datasets, and creates a common configuration named "fixed-parameters". Next, it analyzes the dataset and creates a custom layout according to the following steps: for as many configuration decisions as possible, specifies the dataset properties ("dataset fingerprint") and the design choices ("pipeline fingerprint") in heuristic rule-style to adapt accordingly to the application ("rule-based parameters"); remaining decisions are made directly from the data during training ("empirical parameters").

This method allows for a complete out-of-the-box pipeline that performs segmentation without any specific or manual decision. The pipeline is capable to adapt to several datasets. Training is performed in a 5-fold-Cross-Validation fashion and with three different configurations: 2D, 3D and 3D Cascade. After the training completion, nnU-Net selects empirically the best configuration or ensemble created and post-processing is applied if decided by the empirical parameters. This ensures that the best possible trained model is used for inference.

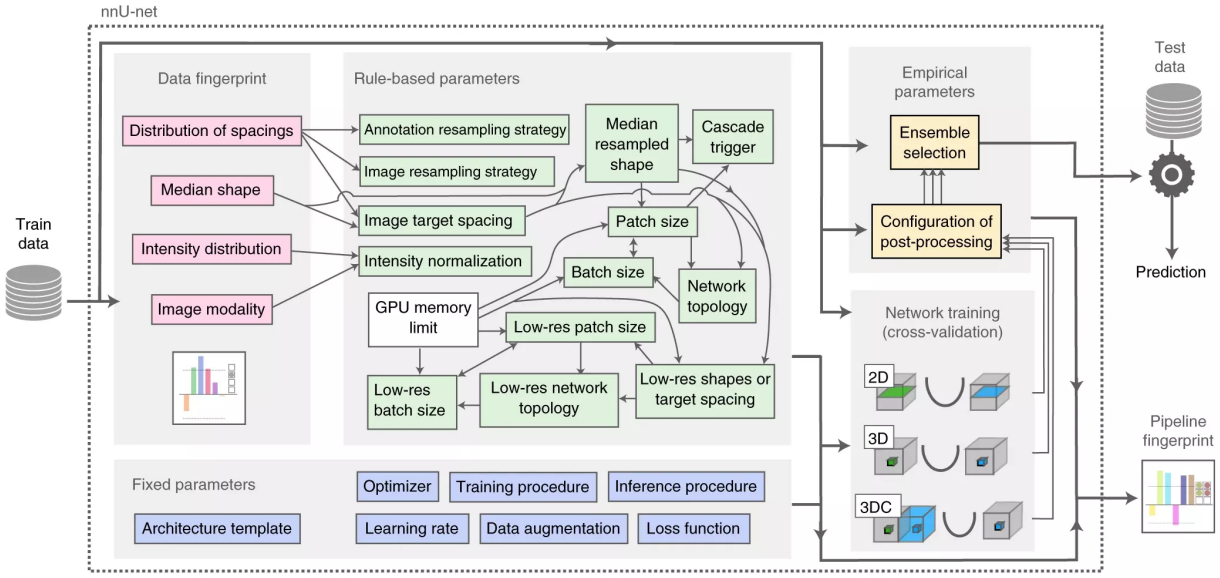


Figure 2.2: Scheme of the nnU-Net architecture and the possible design choices proposed by Isensee et al. [Ise+21]. Dataset properties are shown in pink. Heuristic rules (green) and their interdependencies are shown as thin arrows. Fixed parameters (blue) are pre-established and do not require modifications. Empirical parameters (yellow) are determined during training and control the post-processing. *Note that this image is cropped, for a complete, full-resolution image please refer to the link or Isensee et al. [Ise+21]*

Chapter 3

Materials and Methods

3.1 Dataset

The nnU-Net architecture was trained and evaluated following the same training and validation sets specified in the work of Sukesh et al. [Suk+22]. Table 3.1 shows the structure of both sets and the total number of documents contained in each set. A total of 70 training documents and 56 validation documents were used, as well as their corresponding labels. The total amount of images used increases to 252. Note that, as in the work mentioned above, only track A of the 2019 dataset was used.

3.2 Metrics

Selected metrics to perform the evaluation follow as well the measures specified in the aforementioned work. These include some standard evaluation metrics in the DIBCO challenge. Table 3.2 collects the metrics used for this project: (1) F-Measure (FM), (2) pseudo F-Measure (pFM) [Nti+12], (3) Peak Signal-to-Noise Ratio (PSNR) and (4) Distance Reciprocal Distortion (DRD) [Lu+04]. In addition, two other metrics were evaluated: Negative Rate Measure (NRM) and Missclassification Penalty Measure (MPM). These have been omitted due to their absence in Sukesh et al. [Suk+22]. During nnU-Net training, default pipeline evaluation metric was used, this is, DICE coefficient [Jha+19].

Considering Table 3.2: (1) Precision $= \frac{TP}{TP+FP}$, Recall $= \frac{TP}{TP+FN}$, TP, FP and FN are True Positives, False Positives and False Negatives, respectively; (3) C refers to the maximum fluctuation in the data type, here 8 bits, hence $C = 255$ (RGB image), M and N refer to the width and height of the images, x and y to the current width or height pixel, I

Table 3.1: Training and Evaluation datasets. The split was made following the same specifications used in Sukesh et al. [Suk+22].

(a) Training dataset

Year	No. of Documents
2009	10
2010	10
2011	16
2012	14
2014	10
2016	10
Total	70

(b) Validation dataset

Year	No. of Documents
2013	16
2017	20
2018	10
2019	10
Total	56

Table 3.2: Evaluation metrics used for nnU-Net binarization task.
**range is inverted, lower DRD values mean better binarization performance.*

Metric	Equation	Range (worst to best)
(1) F-Measure (FM)	$\mathrm{FM} = \frac{2 \cdot \mathrm{Recall} \cdot \mathrm{Precision}}{\mathrm{Recall} + \mathrm{Precision}}$	$[0, 1]$
(2) Pseudo F-Measure (pFM)	$\mathrm{pFM} = \frac{2 \cdot \mathrm{pRecall} \cdot \mathrm{pPrecision}}{\mathrm{pRecall} + \mathrm{pPrecision}}$	$[0, 1]$
(3) Peak Signal-to-Nosie Ratio (PSNR)	$\mathrm{PSNR} = 10 \log_{10} \frac{C^2 MN}{\sum_{x=1}^{M} \sum_{y=1}^{N} (I(x,y) - I'(x,y))^2}$	$(0, \infty)$
(4) Distance Reciprocal Distortion (DRD)	$\mathrm{DRD} = \frac{\sum_{k=1}^{S} \mathrm{DRD}_k}{\mathrm{NUBN}}$	$[1, 0]^*$

to the ground truth image and I' to the predicted image. Resulting computations of PSNR are given in dB; (2) pPrecision, pRecall and (4) DRD computations are out of the scope of this project, but detailed explanation can be found in [Nti+12] and [Lu+04], respectively.

3.3 Training and Evaluation

Preprocessing of the data was required in order to transform images into the nnU-Net NIFTI input format. This was accomplished using the tools provided by the nnU-Net framework. Due to the out-of-the-box nature of the analyzed pipeline, no further crop, modification or augmentation procedure was used for the training or evaluation.

Training procedure followed a 5-fold-Cross-Validation, where each fold was trained for 100 epochs only in the 2D image type configuration. It is worth to note that the recommended number of epochs was not followed (1000), due to external time constraints.

Approximate time for completion in each fold was 5 hours and 20 minutes, adding into a total of approximately 1 day, 2 hours and 40 minutes. No additional training was performed for the 3D or 3D Cascade configuration. The GPU used for the training was a NVidia GeForce GTX 1080 8GB.

Inference on the evaluation images was performed using the included tools in the nnU-Net framework. Inference time consumption for all evaluation images was in the range of 1 minute, or approximately 1 second per image. Resulting images were converted back to their original image format for metric computation. The computation of these metrics was performed using an adapted version of the standard DIBCO metrics for the Python language, provided by Peb Ruswono Aryan [Ary]. These computations were further validated using the provided DIBCO tools for each of the available challenges.

Chapter 4

Results

Results of the evaluation between the inference performed by nnU-Net and the ground truth is shown in Table 4.1, as well as the overall best performing algorithm analyzed in Sukesh et al. [Suk+22] and the respective winner of the DIBCO challenge per year. At a first glance, nnU-Net does not perform better than the best overall method previously analyzed, and sometimes (see Table 4.1c/d) even worse than the winner of the challenge in years 2018 and 2019.

There are a few remarkable properties in the performance of nnU-Net. First of all, performance results in the DRD computation are impressive. Not only surpasses the performance of the challenge winners in every year, but it is capable of obtaining better scores than the best overall method analyzed in [Suk+22] (see Table 4.1c/d). Figure 4.1 and Figure 4.2 show some examples of the original images, the binarization predicted by nnU-Net and the ground truth image. These examples may give a hint on the poor FM, pFM and PSNR performance obtained. Although the binarization is clear and without many distortions, this architecture fails to distinguish between the foreground text and background/misplaced text that may exist in different sections of the image. This is usually not the case for background text that lies in the same position as the foreground text. Figure 4.2 shows that pixels of a similar color to those of the handwritten text may trick the architecture, resulting on the large, black borders on the predicted image. The nnU-Net pipeline shows some mild resistance to stains, and performs in an mediocre manner when dealing with low-quality or obfuscated images (see Figure 4.3). Here, the architecture fails to recognize darker stains in front of the text, and also misses some red text at the borders of the image. Nevertheless, the structure of most letters is conserved.

Table 4.1: Results of nnU-Net compared with the best overall method per challenge year in [Suk+22] and the respective winner in DIBCO challenges [Pra+13; Pra+17; Pra+18; Pra+19].

(a) DIBCO2013

Model	PSNR	FM	pFM	DRD
nnU-Net	21.02	92.56	93.49	1.86
DE-GAN	**24.08**	**97.68**	**98.09**	**1.11**
Winner	20.68	92.12	94.19	3.10

(b) DIBCO2017

Model	PSNR	FM	pFM	DRD
nnU-Net	17.73	89.73	91.50	1.76
DP-LinkNet	**22.84**	**97.92**	**97.94**	**0.77**
Winner	18.28	91.04	92.86	3.40

(c) DIBCO2018

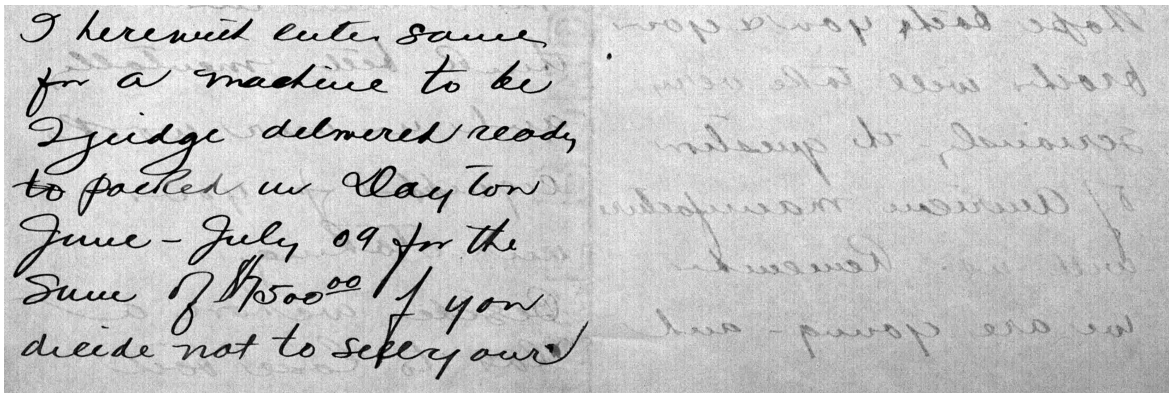
Model	PSNR	FM	pFM	DRD
nnU-Net	13.42	68.45	70.90	**1.89**
2-Stage GAN	**19.93**	**92.40**	**94.90**	2.67
Winner	19.11	88.34	90.24	4.92

(d) DIBCO2019

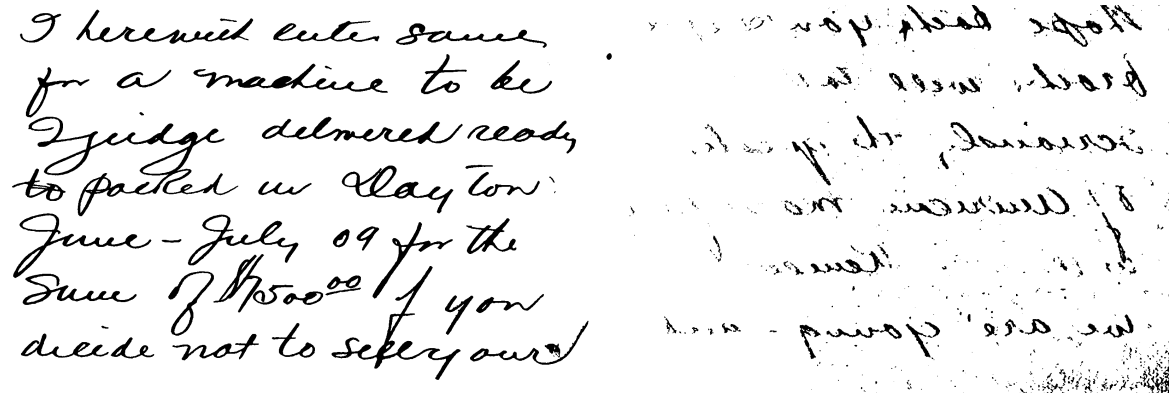
Model	PSNR	FM	pFM	DRD
nnU-Net	14.17	68.86	68.45	**3.57**
SauvolaNet	**15.83**	**72.04**	**71.59**	5.55
Winner	14.48	72.88	72.15	16.24

When compared to all the models analyzed in Sukesh et al. [Suk+22], nnU-Net classifies as 3rd from the last position. Table 4.2 shows the recomputed average statistics (ignoring the img/sec due to the lack of reliable inference processing time), where nnU-Net falls far behind the best performing models in all metrics except for DRD. This states once again the great performance of this pipeline when binarization without distortions is needed.

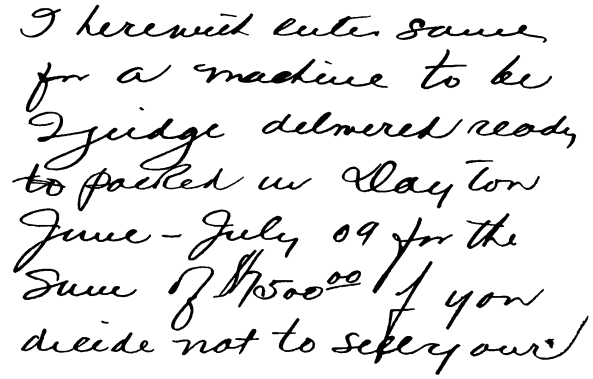
All in all, this poor results may be addressed to the lack of stained, obfuscated or low quality images in the training dataset. In, addition, the recommended time of training was not followed due to external time constraints, and the total number of training epochs was reduced from a 1000 to 100. Figure 4.4 shows the progress of the 1 and 2 fold Cross-Validation steps during training. Although there are some signs of overfitting after reaching 50 epochs (stronger in Figure 4.4b), further training may improve the results obtained in these experiments.



(a) Original image.

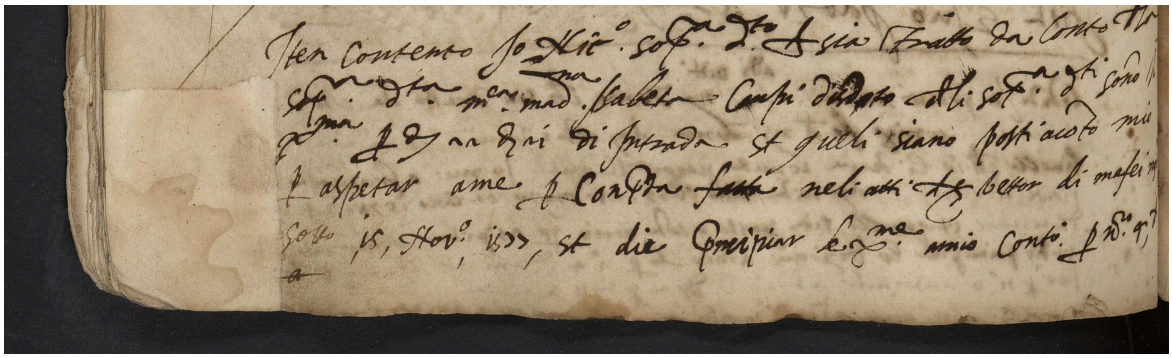


(b) Binarization predicted by nnU-Net.

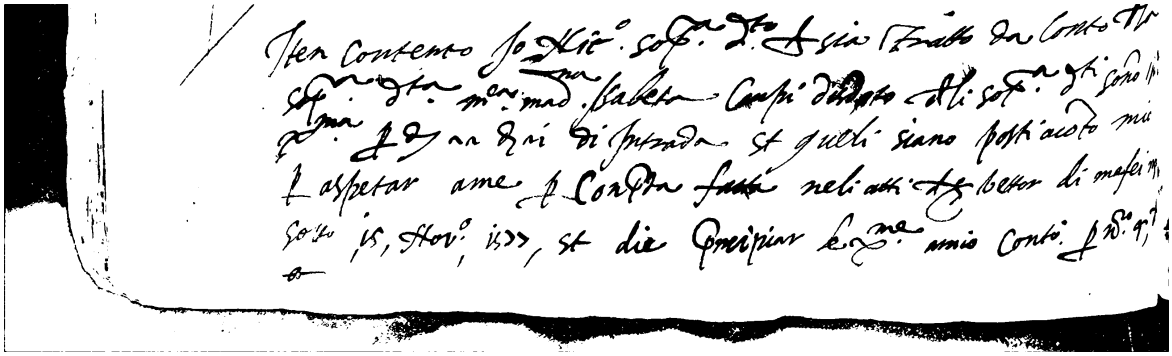


(c) Ground truth binarization of the image.

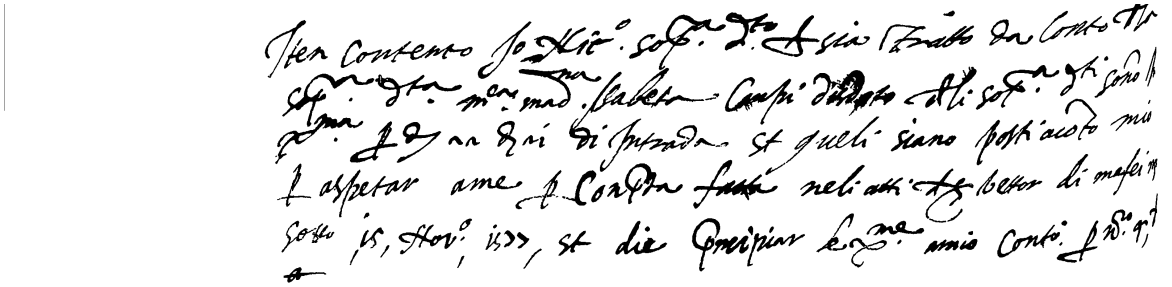
Figure 4.1: Comparison between the original DIBCO2013 [Pra+13] image, the prediction done by nnU-Net and the ground truth image. Although the main text was correctly binarized, the side page was incorrectly inferenced.



(a) Original image.



(b) Binarization predicted by nnU-Net.

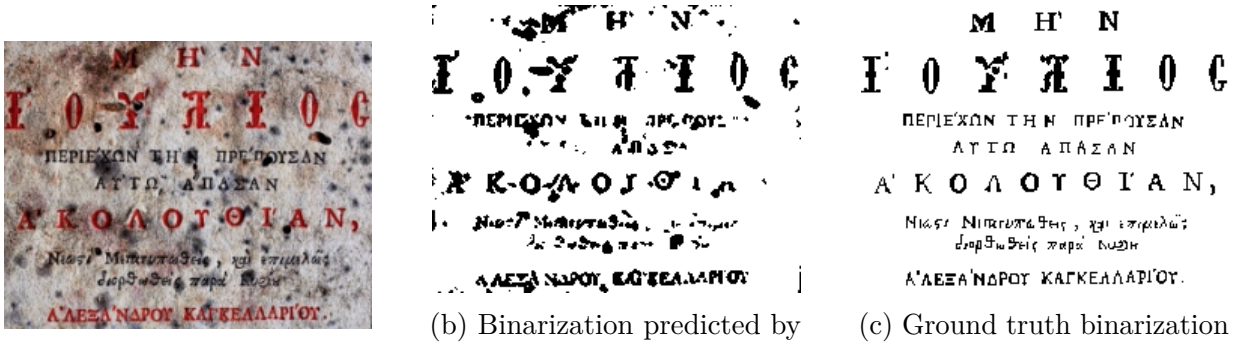


(c) Ground truth binarization of the image.

Figure 4.2: Hard background original DIBCO2018 [Pra+18] image, nnU-Net prediction and ground truth image. It is clearly visible that FM, pFM and PSNR penalized greatly the DIBCO2018 data inference. Nevertheless DRD, shows that the main text is clearly visible, resulting in good performance with this metric (see Table 4.1c).

Table 4.2: Recomputed average metrics for all the analyzed models in Sukesh et al. [Suk+22]

Model	PSNR	FM	pFM	DRD	Avg. Rank
nnU-Net	16.58	79.90	81.08	**2.27**	5.5
DE-GAN	18.37	85.25	87.54	4.64	2.75
Robin (U-Net)	18.24	82.87	84.08	5.95	5.25
DeepOtsu	16.69	79.97	81.89	13.96	6.5
2-Stage GAN	19.07	**87.20**	**88.39**	4.54	**1.5**
DP-LinkNet	**19.10**	83.70	84.36	5.79	3.5
SAE	16.15	79.14	80.77	9.41	7.75
SauvolaNet	18.67	84.91	86.83	4.86	3.25

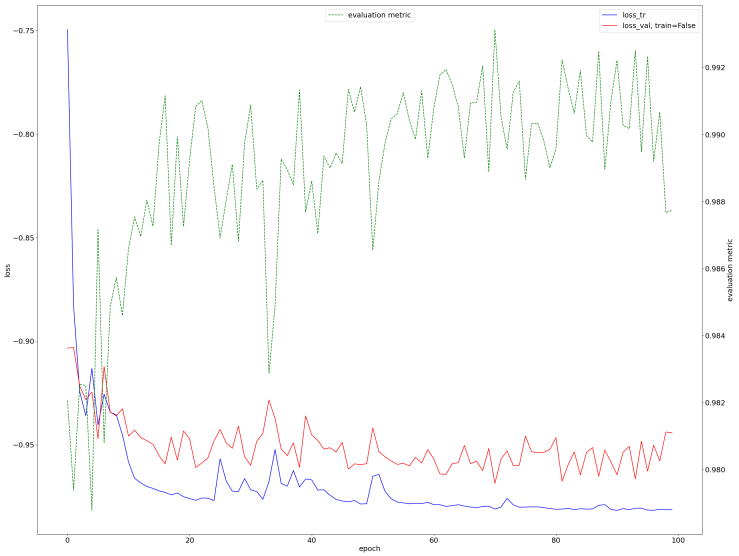


(a) Original image.

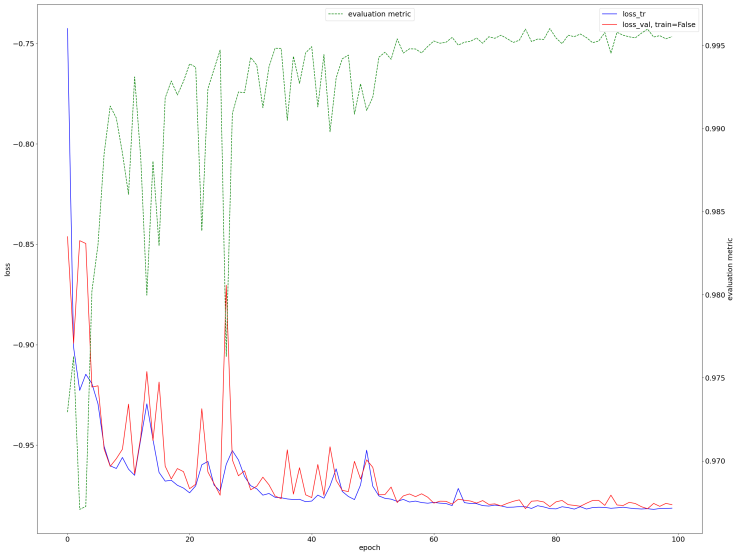
(b) Binarization predicted by nnU-Net.

(c) Ground truth binarization of the image.

Figure 4.3: Example of a difficult and low resolution DIBCO2019 [Pra+19] image, as well as the prediction done by nnU-Net and the ground truth image. It is worth to note that nnU-Net is not easily fooled by stains or low quality images, considering the training data lacks of dirty or stained examples.



(a) Fold 1.



(b) Fold 2.

Figure 4.4: Progress of folds 1 and 2 during nnU-Net training. Fold 2 shows signs of overfitting after 50 epochs, which is less noticeable in fold 1. Key: (green) DICE coefficient [Jha+19]; (blue) training loss; (red) validation loss.

Chapter 5

Conclusion and Future Work

In this project the work of Sukesh et al. [Suk+22] was extended by analizing the performance of the novel, self configuring segmentation pipeline nnU-Net [Ise+21] and thoroughly performed the same experiments as in the above mentioned article. First, the configuration of the nnU-Net pipeline and the training with the 6 selected training datasets was performed, in the default 5-fold-Cross-Validation fashion of nnU-Net. Next, the trained model performed inference on the validation dataset, composed by the remaining 4 datasets. Predictions of nnU-Net model were then evaluated using the same metrics as in the previous work, and the resulting computations compared with the results of other 7 models. Finally, qualitative analysis was performed on the image results, comparing them with the original and ground truth images.

Results clearly show an overall poor to average performance of nnU-Net in binarization problems. Nevertheless, they show impressive performance when the DRD metric comes into the picture, and may be useful when low distortions w.r.t. the original image are a must in certain applications. In addition, the low performance may be caused by the lack of stained, obfuscated or low-resolution images in the training dataset, including images where big black borders or unusual structures appear. This results may have also been influenced by the fact the recommended amount of epochs for training was reduced from 1000 to 100, thus having the model less training time. As well as the time for training, the authors of nnU-Net state that the architecture is more suited for 3D, multi-class specialized datasets, were the model configurations may introduce complexity in the different configurations [Ise+21].

Future work may incorporate more training data with different examples and more difficult situations. Increasing the number of epochs to the recommended ones may also

increase the performance of this model. Finally, during the last steps of this work, nnU-Net was updated from version 1 to version 2, potentially leading to new improvements in the architecture that may further increase the obtained performance.

List of Abbreviations

CNN Convolutional Neural Network

DL Deep Learning

DBI Document Image Binarization

DIBCO Document Image Binarization Contests

FM F-Measure

pFM pseudo F-Measure

PSNR Peak Signal-to-Noise Ratio

DRD Distance Reciprocal Distortion

NRM Negative Rate Measure

MPM Missclassification Penalty Measure

List of Figures

List of Tables

Bibliography

[Ary] P. R. Aryan. H-dibco evaluation metric for binarization algorithms. URL: https://gist.github.com/pebbie/643e28c619efaa2fd30b1595bd5d0e6c (cited on p. 9).

[Chr+17a] V. Christlein, D. Bernecker, F. Hönig, A. Maier, and E. Angelopoulou. Writer identification using gmm supervectors and exemplar-svms. *Pattern Recognition*, 63:258–267, 2017 (cited on p. 1).

[Chr+17b] V. Christlein, M. Gropp, S. Fiel, and A. Maier. Unsupervised feature learning for writer identification and writer retrieval. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 991–997. IEEE, 2017 (cited on p. 1).

[Du+20] G. Du, X. Cao, J. Liang, X. Chen, and Y. Zhan. Medical image segmentation based on u-net: a review. *Journal of Imaging Science and Technology*, 2020 (cited on p. 1).

[Ise+21] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein. Nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, February 2021. ISSN: 1548-7105. DOI: 10.1038/s41592-020-01008-z. URL: https://doi.org/10.1038/s41592-020-01008-z (cited on pp. v, 1, 2, 4, 5, 17).

[Isl+17] N. Islam, Z. Islam, and N. Noor. A survey on optical character recognition system. *arXiv preprint arXiv:1710.05703*, 2017 (cited on p. 1).

[Jha+19] S. Jha, L. H. Son, R. Kumar, I. Priyadarshini, F. Smarandache, and H. V. Long. Neutrosophic image segmentation with dice coefficients. *Measurement*, 134:762–772, 2019. ISSN: 0263-2241. DOI: https://doi.org/10.1016/j.measurement.2018.11.006. URL: https://www.sciencedirect.com/science/article/pii/S0263224118310601 (cited on pp. 7, 16).

[Li+21] D. Li, Y. Wu, and Y. Zhou. Sauvolanet: learning adaptive sauvola network for degraded document binarization. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part IV 16*, pages 538–553. Springer, 2021 (cited on p. 1).

[Lon+15] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015 (cited on p. 3).

[Lu+04] H. Lu, A. Kot, and Y. Shi. Distance-reciprocal distortion measure for binary document images. *IEEE Signal Processing Letters*, 11(2):228–231, 2004. DOI: 10.1109/LSP.2003.821748 (cited on pp. 7, 8).

[Min+22] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos. Image segmentation using deep learning: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3523–3542, 2022. DOI: 10.1109/TPAMI.2021.3059968 (cited on p. 1).

[Nti+12] K. Ntirogiannis, B. Gatos, and I. Pratikakis. Performance evaluation methodology for historical document image binarization. *IEEE Transactions on Image Processing*, 22(2):595–609, 2012 (cited on pp. 7, 8).

[Ots79] N. Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979 (cited on p. 1).

[Pra+09] I. Pratikakis, K. Zagoris, G. Barlas, and B. Gatos. Gatos, b., ntirogiannis, k., pratikakis, i.: icdar 2009 document image binarization contest (dibco 2009). in: 2009 10th international conference on document analysis and recognition, pp. 1375–1382 (2009). In 2009 (cited on p. 2).

[Pra+10] I. Pratikakis, K. Zagoris, G. Barlas, and B. Gatos. Pratikakis, i., gatos, b., ntirogiannis, k.: h-dibco 2010 - handwritten document image binarization competition. in: 2010 12th international conference on frontiers in handwriting recognition, pp. 727–732 (2010). In 2010 (cited on p. 2).

[Pra+11] I. Pratikakis, K. Zagoris, G. Barlas, and B. Gatos. Pratikakis, i., gatos, b., ntirogiannis, k.: icdar 2011 document image binarization contest (dibco 2011). in: 2011 international conference on document analysis and recognition, pp. 1506–1510 (2011). In 2011 (cited on p. 2).

[Pra+12] I. Pratikakis, K. Zagoris, G. Barlas, and B. Gatos. Pratikakis, i., gatos, b., ntirogiannis, k.: icfhr 2012 competition on handwritten document image binarization (h-dibco 2012). in: 2012 international conference on frontiers in handwriting recognition, pp. 817–822 (2012). In 2012 (cited on p. 2).

[Pra+13] I. Pratikakis, K. Zagoris, G. Barlas, and B. Gatos. Pratikakis, i., gatos, b., ntirogiannis, k.: icdar 2013 document image binarization contest (dibco 2013). in: 2013 12th international conference on document analysis and recognition, pp. 1471–1476 (2013). In 2013 (cited on pp. 2, 12, 13).

[Pra+14] I. Pratikakis, K. Zagoris, G. Barlas, and B. Gatos. Ntirogiannis, k., gatos, b., pratikakis, i.: icfhr 2014 competition on handwritten document image binarization (h-dibco 2014). in: 2014 14th international conference on frontiers in handwriting recognition, pp. 809–813 (2014). In 2014 (cited on p. 2).

[Pra+16] I. Pratikakis, K. Zagoris, G. Barlas, and B. Gatos. Pratikakis, i., zagoris, k., barlas, g., gatos, b.: icfhr 2016 handwritten document image binarization contest (h-dibco 2016). in: 2016 15th international conference on frontiers in handwriting recognition (icfhr), pp. 619–623 (2016). In 2016 (cited on p. 2).

[Pra+17] I. Pratikakis, K. Zagoris, G. Barlas, and B. Gatos. Pratikakis, i., zagoris, k., barlas, g., gatos, b.: icdar 2017 competition on document image binarization (dibco 2017). in: 2017 14th iapr international conference on document analysis and recognition (icdar), vol. 01, pp. 1395–1403 (2017). In 2017 (cited on pp. 2, 12).

[Pra+18] I. Pratikakis, K. Zagoris, G. Barlas, and B. Gatos. Pratikakis, i., zagori, k., kaddas, p., gatos, b.: icfhr 2018 competition on handwritten document image binarization (h-dibco 2018). in: 2018 16th international conference on frontiers in handwriting recognition (icfhr), pp. 489–493 (2018). In 2018 (cited on pp. 2, 12, 14).

[Pra+19] I. Pratikakis, K. Zagoris, G. Barlas, and B. Gatos. Pratikakis, i., zagoris, k., karagiannis, x., tsochatzidis, l., mondal, t., marthot-santaniello, i.: icdar 2019 competition on document image binarization (dibco 2019). in: 2019 international conference on document analysis and recognition (icdar), pp. 1547–1556 (2019). In 2019 (cited on pp. 2, 12, 15).

[Ron+15] O. Ronneberger, P. Fischer, and T. Brox. U-net: convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham. Springer International Publishing, 2015. ISBN: 978-3-319-24574-4 (cited on pp. 1, 3, 4).

[Suk+22] R. Sukesh, M. Seuret, A. Nicolaou, M. Mayr, and V. Christlein. A fair evaluation of various deep learning-based document image binarization approaches. In S. Uchida, E. Barney, and V. Eglin, editors, *Document Analysis Systems*, pages 771–785, Cham. Springer International Publishing, 2022. ISBN: 978-3-031-06555-2 (cited on pp. v, 1, 2, 7, 8, 11, 12, 15, 17).

[Ten+20] C. Tensmeyer and T. Martinez. Historical document image binarization: a review. *SN Computer Science*, 1(3):173, May 2020. ISSN: 2661-8907. DOI: 10.1007/s42979-020-00176-1. URL: https://doi.org/10.1007/s42979-020-00176-1 (cited on p. 1).