
http://www.diva-portal.org

Postprint

This is the accepted version of a paper presented at *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*.

Citation for the original published paper:

Westphal, F., Grahn, H., Lavesson, N. (2018)
User Feedback and Uncertainty in User Guided Binarization
In: (pp. 403-410). IEEE Computer Society
https://doi.org/10.1109/ICDMW.2018.00066

N.B. When citing this work, cite the original published paper.

Permanent link to this version:
http://urn.kb.se/resolve?urn=urn:nbn:se:bth-17742

User Feedback and Uncertainty in User Guided Binarization

Florian Westphal, Håkan Grahn, Niklas Lavesson
Department of Computer Science and Engineering
Blekinge Institute of Technology
Karlskrona, Sweden
{*florian.westphal, hakan.grahn,niklas.lavesson*}*@bth.se*

***Abstract*—In a child's development, the child's inherent ability to construct knowledge from new information is as important as explicit instructional guidance. Similarly, mechanisms to produce suitable learning representations, which can be transferred and allow integration of new information are important for artificial learning systems. However, equally important are modes of instructional guidance, which allow the system to learn efficiently. Thus, the challenge for efficient learning is to identify suitable guidance strategies together with suitable learning mechanisms.**

In this paper, we propose guided machine learning as source for suitable guidance strategies, we distinguish between sample selection based and privileged information based strategies and evaluate three sample selection based strategies on a simple transfer learning task. The evaluated strategies are random sample selection, i.e., supervised learning, user based sample selection based on readability, and user based sample selection based on readability and uncertainty. We show that sampling based on readability and uncertainty tends to produce better learning results than the other two strategies. Furthermore, we evaluate the use of the learner's uncertainty for self directed learning and find that effects similar to the Dunning-Kruger effect prevent this use case. The learning task in this study is document image binarization, i.e., the separation of text foreground from page background and the source domain of the transfer are texts written on paper in Latin characters, while the target domain are texts written on palm leaves in Balinese script.

***Index Terms*—guided machine learning, interactive machine learning, image binarization, historical documents**

1. Introduction

When designing systems that learn from experience, it is worthwhile to consider well working learning systems, such as humans and particularly children. This consideration can be used to draw inspiration for approaches to test and directions to pursue. More importantly, comparing effects observed in different learning systems can lead to the discovery of inherent properties of learning systems in general.

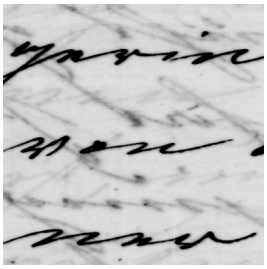
One key aspect of developmental learning is to replicate the ability of humans to transfer previously learned knowledge to new domains or to changed conditions. This ability of humans is captured in the theory of cognitive constructivism of Piaget [1], which views the human learning process as knowledge construction process going through the stages of information selection, organization and integration with previous knowledge [2]. Thus, a large focus of developmental learning lies on designing learning mechanisms, which learn representations from given data. These representations aim to help solving the given task, but must also be transferable and allow integration of new information.

While devising such mechanisms is challenging, we know from human learning and human development that just possessing the ability to construct knowledge is not sufficient for efficient learning. Another crucial factor for efficient learning is the mode of instruction, which needs to fit the learner's proficiency level. For human learning, Clark et al. [3] argue that novices benefit from full and explicit instructional guidance, while experts can benefit from discovery based learning. This notion is reflected in artificial learning systems, for example, in reinforcement learning agents, which essentially perform discovery based learning without prior knowledge of the task, but which are currently unable to learn to play low reward Atari games, such as MONTEZUMA'S REVENGE without expert demonstration [4]. Thus, devising suitable modes of instruction are as important for developmental learning as the design of suitable learning representations for transfer learning.

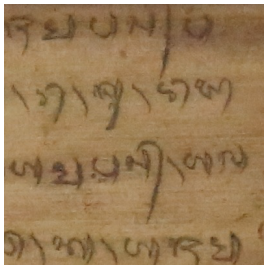
In this paper, we describe guided machine learning (gML) with its different types of guidance as possible framework of thought for different modes of instructional guidance. Furthermore, we discuss and evaluate different forms of guidance for a simple transfer learning task and evaluate the use of the learner's uncertainty for self guided learning in form of active learning [5].

The general learning task used in this paper is document image binarization, i.e., the separation of text foreground from page background. The used algorithm is pre-trained on a source dataset of document images from the Document Image Binarization Contests (DIBCO) [6]–[8] and the Handwritten Document Image Binarization Contests (H-

This work is part of the research project "Scalable resource-efficient systems for big data analytics" funded by the Knowledge Foundation (grant: 20140032) in Sweden.



(a) Image from the H-DIBCO 2016 dataset [12].



(b) Image from the AMADI_LontarSet [13].

Figure 1. Difference in appearance of documents from different collections.

DIBCO) [9]–[12], and the aim is to learn to binarize images of a target dataset from as few labeled samples as possible. The images from the source dataset show handwritten and printed text written in Latin characters on paper, as shown in Figure 1a. The target dataset, on the other hand, is written in Balinese script on palm leaves, as shown in Figure 1b.

The main contributions of this paper are:

- proposal of gML as thought framework for modes of instructional guidance
- evaluation of the learner's uncertainty as source for self guided learning
- evaluation of different instruction strategies
- proposal and evaluation of the use of uncertainty visualization to help user guidance

Our results show that effects similar to the Dunning-Kruger effect [14] interfere with the use of the learner's uncertainty for active learning. Furthermore, we show that user guided learning achieves better binarization results with the same number of labeled training data than normal supervised learning with random sample selection. However, we find that providing the user with a low level explanation of the binarization result, in form of the learner's visualized uncertainty, yields even better binarization results.

In the following, We provide background information about document image binarization and the used learning system in Section 2. Section 3 describes our ideas on gML and Section 4 discusses the estimation and potential pitfalls in estimating the learner's uncertainty. We describe the different strategies for guidance in Section 5 and describe the experiment design in Section 6. The results of the evaluations are presented in Section 7, Section 8 describes related work and Section 9 concludes this paper.

2. Background

2.1. Image Binarization

Image binarization is a document analysis task with the goal to separate text foreground from page background. This is achieved by labeling each pixel in an input image as either foreground or background pixel. Image binarization is a common pre-processing task for other document analysis algorithms, such as word spotting [15] or image transcription [16]. While comparatively simple in contemporary documents, image binarization becomes increasingly difficult in historical documents, due to common image degradations, such as faded ink, stains or ink bleeding through from the other side of the page. Many algorithms have been proposed to address this task, including the algorithms by Otsu [17] and Niblack [18], as well as the more recent deep supervised network approach by Vo et al. [19].

2.2. Recurrent Binarization

The underlying binarization algorithm of our gML system is the recurrent neural network (RNN) based algorithm by Westphal et al. [20]. This algorithm divides a given document image into blocks of 64×64 pixels. Each of these blocks is then read sequentially and individually by four separate Grid long short-term memory (LSTM) [21] network layers, each starting from one of the block's four corners. The output of these layers is combined and processed by another two Grid LSTM layers. The final binarization result is achieved through a weighted sum of the outputs of the two Grid LSTM layers, mapped into the interval between 0 and 1 by the sigmoid function. Thus, each pixel in the processed block receives a continuous value between 0 and 1. To receive the final result, each of these continuous values is rounded to either 0 (foreground) or 1 (background).

3. Guided Machine Learning

In this paper, guided machine learning (gML) refers to a form of interactive machine learning [22], [23], which is concerned only with learning algorithms allowing an external agent to steer the learning process. For defining gML in this paper, we adapt Mitchell's definition of machine learning [24] as follows:

Definition 1. A computer program is said to learn through guidance G from external agent A with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves through actions performed by A with the aim to achieve such an improvement.

Thus, the main focus of gML lies on exploring different guidance mechanisms for steering the learning process. Here, we distinguish between mechanisms based on sample selection, Section 3.1, and mechanisms based on privileged information, Section 3.2. Both directions and their subcategories are described in the following sections.

3.1. Sample Selection

Mechanisms in this category guide the learning process through the selection of training samples. Common selection approaches are *random*, *data focused a priori*, *learner focused a priori* and *learner focused adaptive*.

Random selection. This approach chooses training samples randomly and the only provided guidance are the provided correct labels for the given training samples. This is the standard supervised learning approach.

Data focused a priori selection. This strategy determines the complete training set before starting the training, based on the characteristics of the training data. For example, curriculum learning [25] builds the complete training set by initially choosing easy samples and then adding more and more difficult ones to the training set. This distinction between easy and difficult samples is done based on the data set's characteristics and is independent of the learner.

Learner focused a priori selection. This method determines the complete training set before starting the training, based on the learner's characteristics. Machine teaching [26] employs this strategy by creating an optimal training set based on the knowledge of the target function to be learned and of the learner's learning mechanism.

Learner focused adaptive selection. This approach builds an initial training set, which is iteratively updated, based on the learner's current proficiency. Two approaches following this strategy are iterative machine teaching [27] and Teaching-to-Learn and Learning-to-Teach (TLLT) [28], [29], which aim to generate a new optimal training set after each training iteration based on the learner's current state.

3.2. Privileged Information

On the other hand, privileged information approaches, as described by Vapnik and Izmailov [30], guide the learner by providing additional information about a training sample to the learner. However, this additional information is available only at training time, but not at test time. Here, we distinguish between *a priori selected privileged feature information*, *a priori selected privileged target information* and *adaptively selected privileged information*.

A priori selected privileged feature information. This strategy adds additional features to the training samples regardless of the learner's current performance. Examples for this approach are similarity control and knowledge transfer as described by Vapnik and Izmailov [30], as well as generalized distillation [31]. Another example for this strategy is learning by demonstration as described by Rozo et al. [32]. In their paper, the demonstrated movement trajectory used to teach collaborative behavior to a robot arm can be viewed as privileged information only available at training time.

A priori selected target information. This approach adds richer target information to the training samples regardless of the learner's current performance. One example for this form of guidance is distillation [33], which trains a learner with lower model capacity on the outputs of a pre-trained teacher with higher model capacity.

Adaptively selected privileged information. This method adapts the provided privileged training information based on the learner's current state. Explanatory debugging, as described by Kulesza et al. [34], for example, falls in this category. The system described in their paper illustrates the learner's state for its users by showing them the important words, which led it to classify a message in a certain way. The user can then provide privileged information by adjusting which words are used for the classification and how important those words are for a certain class.

Since we view our learner as novice in the target domain of our transfer learning task, we consider only gML strategies for this task. In this study, we focus on sample selection strategies, in particular on random selection and learner focused adaptive selection. For the latter, samples are selected by a human user by assessing the learner's current proficiency either based solely on its current output or on its output combined with a visualization of its uncertainty. However, we also evaluate the use of the learner's uncertainty for active learning, which is not a gML technique.

4. Estimating Network Uncertainty

The binarization algorithm used in this study labels each image pixel p_i as foreground or background by first assigning to it a continuous value $\hat{y}_i \in [0, 1]$ and then deriving its label l_i as $l_i = \lfloor \hat{y}_i + 0.5 \rfloor$. Since this turns the value 0.5 into the decision boundary between foreground and background, we use the distance of the predicted value $\hat{y}_i$ from this decision boundary as measure for the network's certainty that it assigned a label correctly. Conversely, the distance between $\hat{y}_i$ and the assigned label l_i can be viewed as the network's uncertainty u_i, with:

$$u_i = |\hat{y}_i - l_i| \tag{1}$$

This pixel-level uncertainty estimate can then be used to receive an estimate of the network's labeling uncertainty for any given image by averaging over all pixel uncertainties.

Being able to estimate the network's uncertainty for a given image or a particular area of an image has two potential advantages. First, it should make it possible to identify image areas, which contain many pixels with assigned labels easily changeable through training, and second it should make it possible to identify image areas, which have the highest need for change. The first advantage naturally follows from the definition of our uncertainty measure, i.e., values with high uncertainty lie close to the decision boundary and thus are easily moved across it. However, the second advantage hinges on the relationship between binarization quality and uncertainty, as it assumes a negative correlation between both.

In order to analyze this relationship, we simplify the possible states of one pixel with respect to binarization quality and uncertainty to four distinct states:

- correctly labeled - certain
- correctly labeled - uncertain
- incorrectly labeled - uncertain
- incorrectly labeled - certain

This is a simplification, since there is no sharp border between a certain or uncertain labeling, in contrast to a correct or incorrect labeling. Depending on the state of the

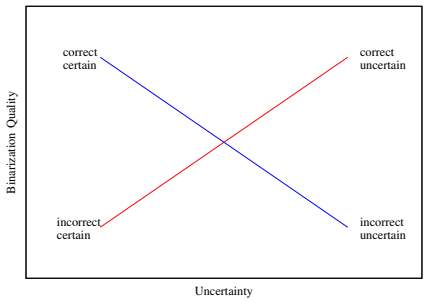


Figure 2. Illustration of the possible relationships between binarization quality and uncertainty, depending on the majority of pixels, with red indicating a positive correlation and blue indicating a negative correlation.

majority of pixels in a binarized image, we can see that the relationship between quality and uncertainty changes. As exemplified by Figure 2, we can expect a negative correlation when the majority of pixels is either labeled correctly with high certainty or if labeled incorrectly with high uncertainty. Conversely, one would expect a positive correlation between binarization quality and uncertainty if the majority of pixels is labeled correctly with high uncertainty or labeled incorrectly with high certainty. This means however that uncertainty can be used to select image areas, which need to be improved, only if the network estimates its uncertainty correctly, i.e., assigns higher uncertainty to incorrectly labeled pixels. But since this estimate relies on the network's ability to label pixels correctly, it should become susceptible to the Dunning-Kruger effect [14], which describes the observation that the lack of ability to perform a certain task may also impede the objective assessment of the correct execution of this task.

5. Learning from User Guidance

In this study, our gML system receives user guidance in form of additional training samples and uses these samples for re-training. Here, it is important to note that the used binarization algorithm binarizes and trains on blocks of 64×64 pixels at once. Therefore, one training sample for this algorithm is one such image block. Additional training samples for re-training can be obtained from users by letting them correct the network's shortcomings, i.e., by providing the correct labels to mislabeled pixels within one such block, and thus producing ground truth data for the block.

Since the used mechanism simply provides ground truth data for training, the question arises: *What is the benefit of learning from user guidance over learning from a ground truth dataset?* The answer to this question is the selection of training samples. While normal training based on a ground truth dataset would require several images to be completely labeled and would then randomly select blocks to train on, in our case of learner focused adaptive selection, blocks to train on are selected by a user and only those blocks need to be labeled. This should lead to less labeled blocks and faster improvements in binarization quality, since a user can easily identify blocks, from which the network can learn the most, i.e., blocks containing many labeling errors. On the other hand, when blocks are chosen randomly, there is no such guarantee. In fact, in this case, blocks containing only background pixels are chosen more often, since document images consist mostly of background pixels.

Additionally to providing users with the network's binarization output to select training samples, it appears reasonable to visualize the network's labeling uncertainty. Since areas of high uncertainty should be easier to change, as described in Section 4, this may help users to select blocks, which will lead to easier improvements or prevent changes in pixels for which the network has found the correct label, but is fairly uncertain about the labeling result. Such a visualization can be achieved easily by quantization of the estimated uncertainty and by mapping each of the quantization levels to a color.

As user guidance focuses on the shortcomings of the current model, a training set selected in this way may overemphasize rare failure cases, while omitting common and currently correctly labeled cases. This, in turn, may lead to worse binarization results after re-training and may increase the need for labeled training samples. One straightforward strategy to avoid this situation is to combine the new user provided training samples with training samples, which were used to build the original model. This combination can help to learn the correct labeling of identified failure cases without forgetting how to label common ones. However, this strategy relies on the fact that the document images on which the original model was trained are similar to the images of the new image collection, which may not always be the case.

6. Experiment Design

6.1. Experiment Setup

All experiments reported in this paper were conducted on a computer with an Intel i7-6700K quad-core CPU @ 4.00 GHz, 32 GB DDR4 RAM and an Nvidia GeForce GTX 980. The implementation of the recurrent binarization algorithm was adapted from the original TensorFlow[1] implementation by Westphal et al. [20] from GitHub[2].

The base model used as starting point for all conducted experiments has been trained on the DIBCO and H-DIBCO datasets, DIBCO 2009 [6], H-DIBCO 2010 [9], H-DIBCO 2012 [10], DIBCO 2013 [8], and H-DIBCO 2014 [11], while the DIBCO 2011 [7] dataset was used for validation and H-DIBCO 2016 [12] for testing. While the DIBCO and H-DIBCO datasets are well suited to build a general model for binarization, due to their inhomogeneity in appearance, they are, for the same reason, not suitable for our experiments. Since the main idea of this study is to re-train a base model to fine-tune it to a particular, homogeneous image collection, a homogeneous dataset is required. Therefore, we chose the AMADI_LontarSet [13], which consists in total

1. https://www.tensorflow.org
2. https://github.com/FlorianWestphal/DAS2018

of 100 images with associated ground truth, and extracted two collections of visually similar images from it. These collections are the *Bangli* set, a collection of 14 images from the Bangli collection, and the *IIIC* set, a collection of 17 visually similar images taken from four different collections. Each of these two sets is split into two subsets of 7 or 8/9 images respectively, which are used for training or testing.

In all conducted experiments, the binarization quality will be assessed using only pseudo F-Measure (F_{ps}), which is an evaluation metric specifically designed to assess binarization quality, since it takes into consideration that certain labeling errors are less severe than others [35]. While other common binarization quality measures, such as F-Measure, peak signal-to-noise ratio (PSNR) and distance reciprocal distortion metric (DRD) [36] were considered, they are omitted here, since they either show the same result as F_{ps}, in case of F-Measure, or overemphasize background labeling errors. This overemphasis can be problematic in initial phases of fine-tuning, as it favors labeling trade-offs, which erase text and thus make the document illegible.

Furthermore, since we are mostly interested in the improvement in binarization performance with respect to the base model, we compute the relative F_{ps} improvement ($\tilde{F}_{ps}^{(t)}$) of a re-trained model t as follows:

$$\tilde{F}_{ps}^{(t)} = \frac{F_{ps}^{(t)} - F_{ps}^{(base)}}{F_{ps}^{(final)} - F_{ps}^{(base)}} \tag{2}$$

Here, $F_{ps}^{(t)}$ denotes the binarization quality achieved by a re-trained model t, while $F_{ps}^{(base)}$ and $F_{ps}^{(final)}$ denote the quality achieved by the base model and a model trained to convergence on the target collection's complete training set respectively. $F_{ps}^{(final)}$ is trained in this way to allow an estimation of the potential room for improvement starting from the base model's performance and to put the binarization quality achieved by the evaluated guidance strategies into perspective.

6.2. Evaluating Uncertainty and Quality

In order to analyze the relationship between network uncertainty and binarization quality, we compute a model's uncertainty on all images of a given test set and measure then the binarization performance of this model. Based on these values, we can compute Pearson's correlation coefficient ρ to measure the linear correlation between uncertainty and binarization quality.

We compute this correlation coefficient for the seven used DIBCO and H-DIBCO datasets by training a model on all but two of these datasets, and use one of the held out datasets for validation and the other for testing. We then combine the results for network uncertainty and binarization quality on the test dataset from all these seven models by standardizing them, and compute their overall correlation coefficient. Similarly, we combine the two sub-collections of the Bangli and IIIC dataset by building two models per dataset, standardizing their results and then combining them.

While there is only one final model in case of the DIBCO and H-DIBCO datasets, for the Bangli and IIIC dataset, we look at the correlation coefficient for the base model, as well as for the final model, i.e., the model produced from the base model after training to convergence using the respective training dataset.

6.3. Evaluating Guidance Strategies

In this paper, we evaluate 5 different guidance strategies for selecting image blocks to train on. These strategies are supervised learning, i.e., random selection (f_{rnd}), selection based on readability (f_r), selection based on uncertainty and readability (f_{ur}), as well as a combination of blocks selected by f_r or f_{ur} with blocks from the source training set, denoted as f_{rc} and f_{urc} respectively. For f_r, we had one user selecting an equal number of image blocks from all training set images, adding up to a total of 512 blocks per training set. These blocks were selected based on the user's perception of which mislabeled blocks affected the readability most negatively. The same number of image blocks was selected by the same user for f_{ur} with the key difference that the user had access to the network's visualized uncertainty and could therefore make use of this knowledge during selection.

Since the overall aim of this study is to reduce the number of labeled image blocks, we analyze the impact different selection strategies have on the binarization performance when taking subsets from the 512 selected training blocks of size 32, 64, 128, 256 and 512. Each of these five subsets is then used to re-train the base model for 5 training epochs with a batch size of 8 blocks per batch. Thus, the weights of the five produced models are updated $5 \cdot (4, 8, 16, 32, 64)$ times respectively. The produced models are then used to assess the strategy's impact on the binarization performance. While f_{rc} and f_{urc} use the same image blocks as f_r and f_{ur} respectively, their number of training blocks is always twice as large, since they combine, e.g., 32 blocks from f_r with another 32 blocks from the base model's training set.

7. Results and Analysis

7.1. Uncertainty and Binarization Quality

Table 1 shows the correlation coefficients and binarization quality measured in F_{ps} for the DIBCO/H-DIBCO, Bangli and IIIC datasets. In this table, (base) denotes the base model's binarization quality on the respective dataset, while (final) denotes the quality of the model trained on all available training samples for the respective dataset until convergence. In case of the DIBCO/H-DIBCO datasets, we can see the negative correlation between uncertainty and binarization quality, one would expect from a model proficient in labeling images from a certain dataset. This is the case, since the majority of pixels is labeled correctly with high certainty, while mislabeled pixels have a high uncertainty (cf. Figures 3a and 3b). Thus, uncertainty could be used to identify blocks, which should be used for training.

TABLE 1. CORRELATION COEFFICIENTS, AND AVERAGE F_{ps} AND UNCERTAINTY RESULTS FOR THE ANALYZED DATASETS AND MODELS.

	ρ	F_{ps}	u_{avg}
DIBCO/H-DIBCO	-0.416	90.960	0.013
Bangli (base)	0.158	27.930	0.045
Bangli (final)	0.157	54.797	0.044
IIIC (base)	0.813	51.003	0.025
IIIC (final)	0.444	58.160	0.052

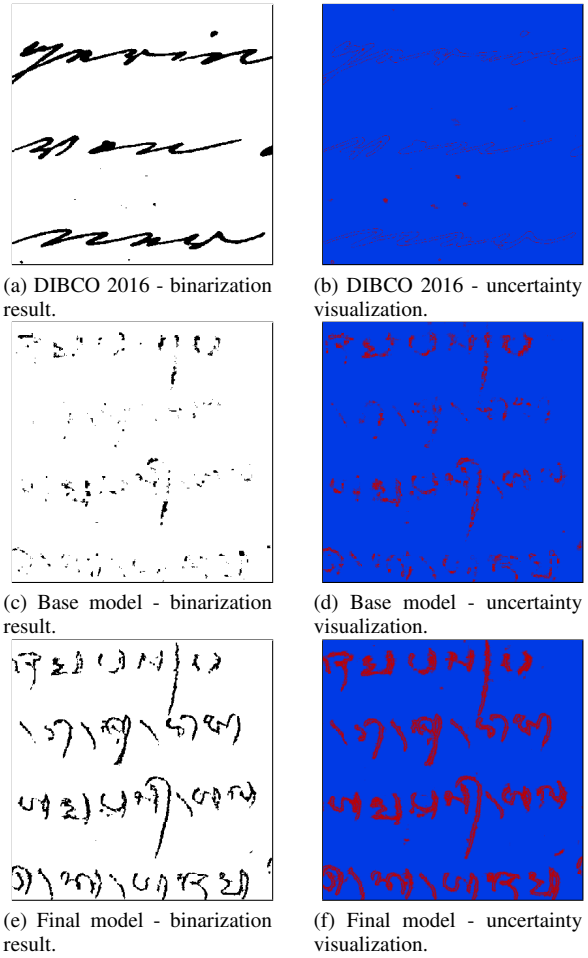
However, for the base model applied to the Bangli and IIIC dataset, we observe a positive correlation with relatively low binarization performance. This is caused by a majority of pixels being labeled either incorrectly with high certainty or correctly with high uncertainty. Figures 3c and 3d exemplify this by showing that, in case of the Bangli dataset, the base model has the tendency to mislabel foreground pixels with high confidence, explaining the low binarization performance. This resembles the Dunning-Kruger effect, since the network's lacking proficiency in finding the correct foreground labels also impedes its uncertainty estimation.

While one would expect a better uncertainty estimate and thus a negative correlation between uncertainty and binarization quality after training to convergence on a collection's complete training set, this is not the case for the Bangli and IIIC dataset, as shown in Table 1. This observation is especially striking for the Bangli dataset, as we see no change in neither correlation nor average uncertainty. However, when looking at the example images in Figures 3e and 3f, we can see that the labeling performance changes from labeling the majority of pixels incorrectly with high certainty towards labeling the majority of pixels correctly with high uncertainty. Thus, the correlation remains positive, even though the binarization performance increases. While not visible in the shown examples, we can observe that the overall average uncertainty stays the same between base and final model, because the areas of increased uncertainty are re-distributed through training, from the dark background regions of the image towards the foreground text.

Lastly, we observe for the Bangli and IIIC dataset that foreground pixels are assigned a high uncertainty after training, compared to the DIBCO and H-DIBCO datasets (cf. Fig. 3b and 3f). This appears to be an artifact caused by the two ground truth images per image in the AMADI_LontarSet, which do not always agree with each other or with the actual source image in such a form that ground truth strokes are sometimes significantly thinner than the actual character strokes or lie outside the area of the actual character. Therefore, it is reasonable to assume that the network does not reach the same level of certainty in labeling foreground pixels than when being trained on one reliable ground truth dataset.

7.2. Guidance Strategies

Figure 4 shows the relative improvements in F_{ps} for all images of the Bangli dataset, which were achieved by models trained according to the 5 evaluated strategies



(a) DIBCO 2016 - binarization result.
(b) DIBCO 2016 - uncertainty visualization.
(c) Base model - binarization result.
(d) Base model - uncertainty visualization.
(e) Final model - binarization result.
(f) Final model - uncertainty visualization.

Figure 3. Binarization result and uncertainty estimate for different models on the image from Fig. 1a and 1b. Red identifies high uncertainty pixels, while low uncertainty pixels are shown in blue.

on 32, 64, 128, 256 or 512 training blocks. In order to find statistically significant differences between the different strategies and numbers of used training blocks, we perform the Friedman test between all strategies and training block numbers. This test shows a statistically significant difference between strategies and number of blocks used for training at the $p < 0.05$ level. Therefore, we perform pairwise Wilcoxon rank-sum tests with Holm correction to find differences between the strategies.

The most interesting finding from these pairwise tests is that f_r and f_{ur} perform consistently, statistically significantly better than f_{rnd} trained on 32 blocks, given that f_r and f_{ur} trained on more than 32 blocks. While this appears like a low bar, it is not achieved by any of the other strategies, which indicates that these two strategies produce a better binarization performance with less training samples. Furthermore, it can be seen from Figure 4 that f_{ur} produces models, which perform more consistently and better than f_r. The main reason for this may be that including considerations about the network uncertainty in the selection process lead the user to create more representative training

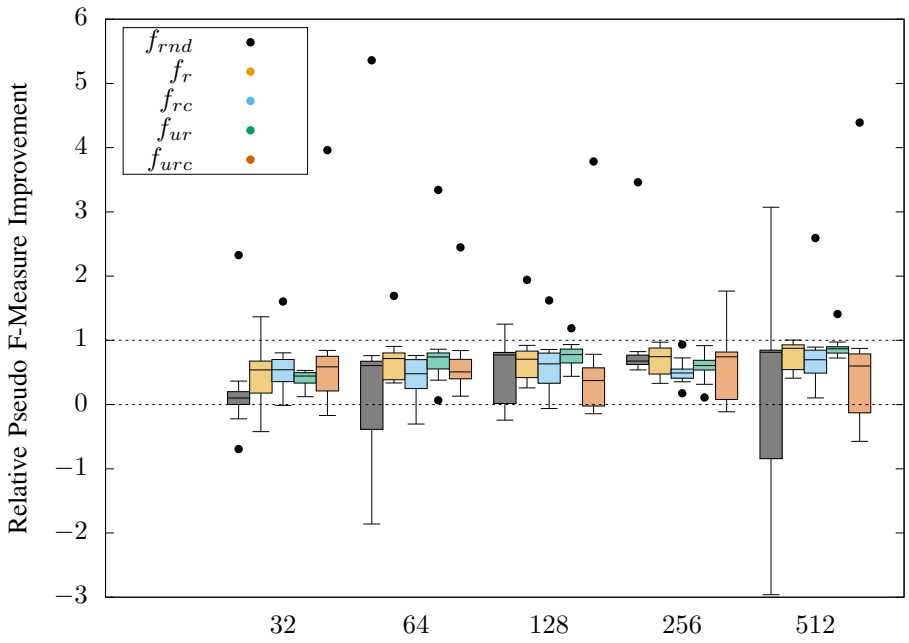


Figure 4. Relative improvements in F_{ps} for the different feedback strategies and different numbers of feedback blocks for the Bangli dataset. The strategies depicted in this plot are random selection (f_{rnd}), selection based on readability (f_r), selection based on uncertainty and readability (f_{ur}), and the respective combined approaches (f_{rc}) and (f_{urc}). Boxes illustrate the first, second and third quartile; Whiskers indicate the lowest and the highest value within the 1.5 interquartile range; Dots mark outliers outside this range.

sets, which result in better models.

Additionally, one can observe from Figure 4 that f_{rc} and f_{urc} tend to perform worse in relation to the other strategies the more blocks are used for training. This appears to be the case due to the dissimilarity between the DIBCO/H-DIBCO datasets and the Bangli dataset, which prevents a proper fine-tuning to the Bangli dataset if too many DIBCO/H-DIBCO samples are used for re-training.

The results for the IIIC dataset are omitted here, since they are not sufficient to draw any conclusions about the evaluated 5 strategies. This may be the case, since the base model performs well on the IIIC dataset and training does not result in a large improvement over this initial performance, as shown in Table 1. Therefore, the potential room for improvement may not be large enough to observe clear differences between the different strategies.

8. Related Work

Within document analysis, several approaches have been proposed, which incorporate user feedback to achieve better solutions to their respective analysis task. Examples for such systems are the document retrieval system by Rusiñol and Lladós [37] or the document recognition system by Carton et al. [38]. However, the work most closely related to our own is the work by Huang et al. [39]. Their proposed approach uses pixels labeled by the user to learn to remove ink bleeding through from the other side of the page. The works by Huang et al., and Rusiñol and Lladós differ from our work since their aim is to improve their approach's performance only on the current sample, while we aim to improve the overall performance of our model. While Carton et al.'s goal is similar to ours in this respect, their approach is mostly to provide aid to the user to define rules for their rule based document recognition system.

9. Conclusion

In this paper, we have proposed gML as source for efficient modes of instruction for systems within developmental learning. Further, we have compared different sample selection based gML strategies with each other on a simple transfer learning task. The evaluated strategies were random selection (f_{rnd}), i.e., supervised learning, and learner focused adaptive selection with either user selected samples based on readability (f_r) or user selected samples based on readability and uncertainty (f_{ur}). We show that in this case, f_{ur} tends to perform better than the other two strategies. The reason for this tendency may be that the combination of selection based on readability with selection based on visualized uncertainty results in more representative training sets, i.e., training sets, which do not overemphasize text regions, since background regions are selected as well, due to their high uncertainty. However, further work is required to confirm this tendency for other tasks or datasets.

Apart from the evaluation of different guidance strategies, we have shown that the relationship between network uncertainty and binarization performance depends on the

network's labeling proficiency. While well trained networks show a negative correlation between uncertainty and binarization quality, networks trained on images from different collection tend towards a positive correlation between these two measures. We have argued that this positive correlation is caused by the network's overconfidence in its incorrect label assignments, which is caused by the network's low proficiency - a problem reminding of the Dunning-Kruger effect. This effect prevents the use of network uncertainty as single guide for training sample selection.

References

[1] J. Piaget, *Science of education and the psychology of the child.* Orion, 1970.

[2] R. E. Mayer, "Constructivism as a theory of learning versus constructivism as a prescription for instruction," *Constructivist instruction: Success or failure*, pp. 184–200, 2009.

[3] R. Clark, P. A. Kirschner, and J. Sweller, "Putting students on the path to learning: The case for fully guided instruction," *American Educator*, 2012.

[4] Y. Aytar, T. Pfaff, D. Budden, T. L. Paine, Z. Wang, and N. de Freitas, "Playing hard exploration games by watching youtube," *arXiv preprint arXiv:1805.11592*, 2018.

[5] B. Settles, "Active learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 6, no. 1, pp. 1–114, 2012.

[6] B. Gatos, K. Ntirogiannis, and I. Pratikakis, "DIBCO 2009: document image binarization contest," *Int. J. Document Anal. and Recognition*, vol. 14, no. 1, pp. 35–44, 2011.

[7] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "ICDAR 2011 Document Image Binarization Contest (DIBCO 2011)," in *Int. Conf. Document Anal. and Recognition*, 2011, pp. 1506–1510.

[8] ——, "ICDAR 2013 Document Image Binarization Contest (DIBCO 2013)," in *Int. Conf. Document Anal. and Recognition*, 2013, pp. 1471–1476.

[9] ——, "H-DIBCO 2010-Handwritten Document Image Binarization Competition," in *Int. Conf. Frontiers in Handwriting Recognition*, 2010, pp. 727–732.

[10] ——, "ICFHR 2012 competition on handwritten document image binarization (H-DIBCO 2012)," in *Int. Conf. Frontiers in Handwriting Recognition*, 2012, pp. 817–822.

[11] K. Ntirogiannis, B. Gatos, and I. Pratikakis, "ICFHR2014 Competition on Handwritten Document Image Binarization (H-DIBCO 2014)," in *Int. Conf. Frontiers in Handwriting Recognition*, 2014, pp. 809–813.

[12] I. Pratikakis, K. Zagoris, G. Barlas, and B. Gatos, "ICFHR2016 Handwritten Document Image Binarization Contest (H-DIBCO 2016)," in *Int. Conf. Frontiers in Handwriting Recognition*, 2016, pp. 619–623.

[13] M. W. A. Kesiman, J. C. Burie, G. N. M. A. Wibawantara, I. M. G. Sunarya, and J. M. Ogier, "AMADI_lontarset: The first handwritten balinese palm leaf manuscripts dataset," in *15th Int. Conf. Frontiers in Handwriting Recognition*, 2016, pp. 168–173.

[14] J. Kruger and D. Dunning, "Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments." *Journal of personality and social psychology*, vol. 77, no. 6, p. 1121, 1999.

[15] D. Doermann, "The indexing and retrieval of document images: A survey," *Computer Vision and Image Understanding*, vol. 70, no. 3, pp. 287–298, 1998.

[16] A. Graves and J. Schmidhuber, "Offline handwriting recognition with multidimensional recurrent neural networks," in *Advances in neural information processing systems*, 2009, pp. 545–552.

[17] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1979.

[18] W. Niblack, *An Introduction to Digital Image Processing*. Prentice-Hall, Englewood Cliffs, 1986, pp. 115–116.

[19] Q. N. Vo, S. H. Kim, H. J. Yang, and G. Lee, "Binarization of degraded document images based on hierarchical deep supervised network," *Pattern Recognition*, vol. 74, pp. 568–586, 2018.

[20] F. Westphal, N. Lavesson, and H. Grahn, "Document image binarization using recurrent neural networks," in *13th IAPR Int. Workshop on Document Anal. Systems*. IEEE, 2018, pp. 263–268.

[21] N. Kalchbrenner, I. Danihelka, and A. Graves, "Grid long short-term memory," *arXiv preprint arXiv:1507.01526*, 2015.

[22] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza, "Power to the people: The role of humans in interactive machine learning," *AI Magazine*, vol. 35, no. 4, pp. 105–120, 2014.

[23] A. Holzinger, "Interactive machine learning for health informatics: when do we need the human-in-the-loop?" *Brain Informatics*, vol. 3, no. 2, pp. 119–131, 2016.

[24] T. M. Mitchell, *Machine learning*. McGraw-Hill, 1997.

[25] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. 26th Annu. Int. Conf. Machine Learning*. ACM, 2009, pp. 41–48.

[26] X. Zhu, "Machine teaching: An inverse problem to machine learning and an approach toward optimal education," in *29th AAAI Conference on Artificial Intelligence*, 2015.

[27] W. Liu, B. Dai, A. Humayun, C. Tay, C. Yu, L. B. Smith, J. M. Rehg, and L. Song, "Iterative machine teaching," in *International Conference on Machine Learning*, 2017, pp. 2149–2158.

[28] C. Gong, D. Tao, J. Yang, and W. Liu, "Teaching-to-learn and learning-to-teach for multi-label propagation," in *30th AAAI Conference on Artificial Intelligence*, 2016.

[29] C. Gong, D. Tao, W. Liu, L. Liu, and J. Yang, "Label propagation via teaching-to-learn and learning-to-teach," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 28, no. 6, pp. 1452–1465, 2017.

[30] V. Vapnik and R. Izmailov, "Learning using privileged information: Similarity control and knowledge transfer," *Journal of Machine Learning Research*, vol. 16, pp. 2023–2049, 2015.

[31] D. Lopez-Paz, L. Bottou, B. Schölkopf, and V. Vapnik, "Unifying distillation and privileged information," *arXiv:1511.03643 [cs, stat]*, 2015.

[32] L. Rozo, J. Silvério, S. Calinon, and D. G. Caldwell, "Exploiting interaction dynamics for learning collaborative robot behaviors," *Proc. of the Interactive Machine Learning Workshop, co-located with the 2016 International Joint Conference on Artificial Intelligence*, 2016.

[33] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv:1503.02531 [cs, stat]*, 2015.

[34] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf, "Principles of explanatory debugging to personalize interactive machine learning," in *Int. Conf. Intelligent User Interfaces*, 2015, pp. 126–137.

[35] K. Ntirogiannis, B. Gatos, and I. Pratikakis, "Performance Evaluation Methodology for Historical Document Image Binarization," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 595–609, Feb 2013.

[36] H. Lu, A. C. Kot, and Y. Q. Shi, "Distance-reciprocal distortion measure for binary document images," *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 228–231, 2004.

[37] M. Rusiñol and J. Lladós, "Boosting the handwritten word spotting experience by including the user in the loop," *Pattern Recognition*, vol. 47, no. 3, pp. 1063–1072, 2014.

[38] C. Carton, A. Lemaitre, and B. Coüasnon, "Eyes wide open: an interactive learning method for the design of rule-based systems," *Int. J. Document Anal. and Recognition*, vol. 20, no. 2, pp. 91–103, 2017.

[39] Y. Huang, M. S. Brown, and D. Xu, "A framework for reducing ink-bleed in old documents," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–7.