Transfer Learning for Handwriting Recognition on Historical Documents

Adeline Granet, Emmanuel Morin, Harold Mouchère, Solen Quiniou and Christian Viard-Gaudin
LS2N, UMR CNRS 6004, Université de Nantes, France

Keywords: Handwriting Recognition, Historical Document, Transfer Learning, Deep Neural Network, Unlabeled Data.

Abstract: In this work, we investigate handwriting recognition on new historical handwritten documents using transfer learning. Establishing a manual ground-truth of a new collection of handwritten documents is time consuming but needed to train and to test recognition systems. We want to implement a recognition system without performing this annotation step. Our research deals with transfer learning from heterogeneous datasets with a ground-truth and sharing common properties with a new dataset that has no ground-truth. The main difficulties of transfer learning lie in changes in the writing style, the vocabulary, and the named entities over centuries and datasets. In our experiment, we show how a CNN-BLSTM-CTC neural network behaves, for the task of transcribing handwritten titles of plays of the Italian Comedy, when trained on combinations of various datasets such as RIMES, Georges Washington, and Los Esposalles. We show that the choice of the training datasets and the merging methods are determinant to the results of the transfer learning task.

1 INTRODUCTION

Historical documents are more and more digitized to preserve them and to ease their accessibility and diffusion. Thus, information retrieval within historical documents is a real challenge. Moreover, the quantity of images is so large that manual information mining remains a time-consuming task. Over the last decade, historical data have become a principal target for classification (Cloppet et al., 2016), line detection (Murdock et al., 2015), and keyword spotting (Puigcerver et al., 2015). Standard end-to-end text recognition systems consist of three steps (Fischer et al., 2009): manual labelling of data to create a ground truth; specific pre-processing operations such as denoising documents and segmenting them into blocks, lines, or words; training of a dedicated recognizer using the alignment between text images and manual labels.

In (Lladós et al., 2012), the authors study the problem of handwriting recognition (HWR) without training data for historical documents. Using a non dedicated dataset is ill-advised because there may be several problems caused by significant differences in terms of period and geographical area, which often affect the script style. Finding a training dataset for keyword spotting or handwriting recognition, meeting the desired distinctive characteristics, is a complicated task. However, some studies attempt to use modern data spotting systems for historical documents, as (Frinken et al., 2010) which mixes different resources to transcribe another resource based on its target vocabulary: this approach is called *Transfer Learning*. We are especially interested in *Transductive Transfer Learning* which focuses on domain adaptation (Pan and Yang, 2010) by using various source and target data for the same given task. We want to use this method and to push it even further by multiplying the number of annotated data used as sources and by adding parameter transfer.

In this paper, we are studying a new resource (financial records of the Italian Comedy) using a minimum amount of information and without a ground-truth. This prevents the direct use of traditional methods of HWR (see section 2). That is why we want to build a recognition system able to transfer knowledge on unknown data, without annotating more data. To overcome the lack of data, the *Transductive Transfer Learning* seems to be a good alternative. The chosen recognition system is a BLSTM-CTC system (Bidirectional Long Short-Term Memory and Connectionist Temporal Classification) including a Full Convolution Network, among the current state-of-the-art systems; this avoids abstracting the specific extraction of features. Three datasets are used during the training step of the network, each dataset sharing at least one feature with our Italian Comedy data.

The rest of this paper is organized as follows. We present the state of the art on handwriting recognition systems in Section 2. Then, we describe our HWR system, its structure, and its relevant post-

Granet, A., Morin, E., Mouchère, H., Quiniou, S. and Viard-Gaudin, C.
Transfer Learning for Handwriting Recognition on Historical Documents.
DOI: 10.5220/0006598804320439
In *Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2018)*, pages 432-439
ISBN: 978-989-758-276-9

preprocessing steps, in Section 3. The dataset we use is presented in Section 4 and we report our experiments and results in Sections 5 and 6.

2 STANDARD HWR SYSTEMS

At the end of the 90's, Hidden Markov Models (HMMs) had become a reference due to their ability to learn sequentially. Moreover, they could integrate knowledge in the form of lexicons and language models to better label sequences (Bunke et al., 1995). Rapidly, they were strengthened by neural networks and hybrid neuro-markovian systems further improved the local and global representation of characters (Koerich et al., 2002).

At the same time, neural networks evolved with new types of neurons, namely recurrent neurons. Contrary to a simple neuron, a recurrent neuron allows a connection to itself. Thus, recurrent neural networks (RNN) store more information from all inputs during training (Senior and Robinson, 1998). Then, the Long Short-Term Memory (LSTM) block (Hochreiter and Schmidhuber, 1997) appeared to solve the problem of the vanishing gradient. This block allows the training of the network to converge.

Recently, multi-dimensional neural networks with LSTM (MDLSTM) have outperformed traditional networks. The first one was a bidirectional recurrent network with LSTM (BLSTM); it became a reference for HWR systems (Fischer et al., 2009). Nowadays, multidimensional recurrent neural networks won competitions such as (Grosicki and El Abed, 2009). In (Graves and Schmidhuber, 2009), the multidimensional part of MDLSTM is made of four parallel layers across each direction on raw images. All the context can thus be used without any restriction.

In HWR, all the previously introduced networks cannot automatically align the input sequence with its labels. So, (Graves, 2012) presented a connectionist temporal classification (CTC) layer which computes a sequence of labels to avoid the segmentation of the input sequence into characters or words. Another important step in HWR is feature extraction. Although traditional methods such as HOG (Terasawa and Tanaka, 2009) have proven their efficiency, new methods integrating convolutional neural networks (CNN) have begun to replace them (Suryani et al., 2016). Other fields like Visual Recognition and Description also use methods based on CNN and LSTM (Donahue et al., 2015). We can distinguish two different approaches: the first one is a simple CNN and includes at least one layer of full connected neurons at the end of the network, and the second one mainly uses convolution and max pooling layers.

Multilingual systems were developed in parallel of MDLSTM. Some use identical configurations on independent training datasets (Voigtlaender et al., 2016), others dedicate one specific layer for each language and for each task in a recurrent neural network (Moysset et al., 2014) but few systems are trained on multilingual datasets at the same time (Kozielski et al., 2014). Regarding monolingual or multilingual HWR systems, it is common to use *n*-gram language models at the word or character level, and a dictionary closed on the training set to improve the decoding step. In (Oprean et al., 2013), the authors use Wikipedia to create a dynamic dictionary for each word detected as out-of-vocabulary.

3 FCN-BLSTM-CTC RECOGNITION SYSTEM

In our HWR systems, the concept of *Transductive Transfer Learning* is performed through the training and the validation of the weights of the neural network on a set of three datasets. Then, the saved weights are used either to perform tests directly on the new dataset or to perform fine-tuning by initializing the new network. Our neural networks are made up of two parts: feature extraction and handwriting recognition.

The feature extraction part is directly integrated in the HWR system with a fully convolutional neural network (FCN). A graphical representation of the two architectures that we are using in our experiments is presented in Figure 1. The neural network takes an input image with a fixed-height of 120 pixels and a variable width denoted t. The common part of both networks is composed of three layers of convolution with a kernel size of 5x5 and same-padding (corresponding to the gray part of Figure 1). Then, the first network, called *CNN_32*, has 3 layers of convolution with 32 filters, while the second one, called *CNN_128*, has 3 layers of convolution with a filter number which doubles: 64 and 128 filters. Each of the convolutional layers is followed by several max-pooling layers. Finally, the last shape obtained has dimensions $32\text{x}1\text{x}\frac{t}{20}$ for *CNN_32*, and $128\text{x}1\text{x}\frac{t}{20}$ for *CNN_128*. This first part is directly connected to the second part of the network, aka the handwriting recognition system.

The second part of our neural networks respects the initial structure of the BLSTM neural network proposed in (Graves, 2012). The network is composed of two hidden layers, forward and backward. Both of them are made up of 100 LSTM blocks. Through the training, these layers are independent, one using the information following the time, and the

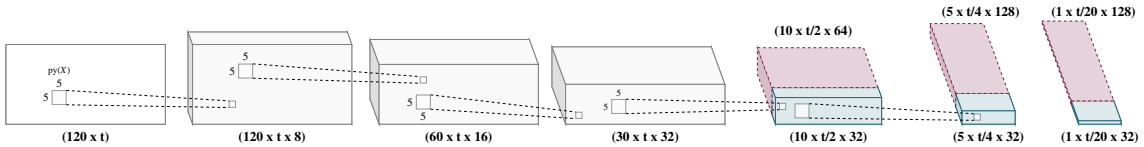


Figure 1: Graphical representation of the two architectures of the convolutional neural network part used for feature extraction. The blue version is called CNN_32, and the pink one is called CNN_128.

other from the future to the past. LSTM blocks control the influence of long-term information across the network; it is interesting for long images such as lines of text. Then, the weighted sum of both hidden layers is provided to the output layer built with 75 *softmax* neurons. This corresponds to 52 lower and upper characters, 10 digits, and some punctuation symbols. At each time step, one output neuron represents one character and an additional neuron acts as a "joker", called "blank" label. Finally, the CTC is applied to the output in order to label the sequence. The provided output can be decoded by several algorithms, such as the *Token Passing Algorithm*, or the *Prefix Search Decoding* that can include a language model. In our experiments, the *Best Path Decoding* is used: at each time step, the most active node is selected, which finally gives the most probable path (Graves, 2012). To obtain the final sequence, all consecutive character labels are deleted except the first one, as well as blank labels. With this method, we stay in the framework where we have no *prior* knowledge on the data.

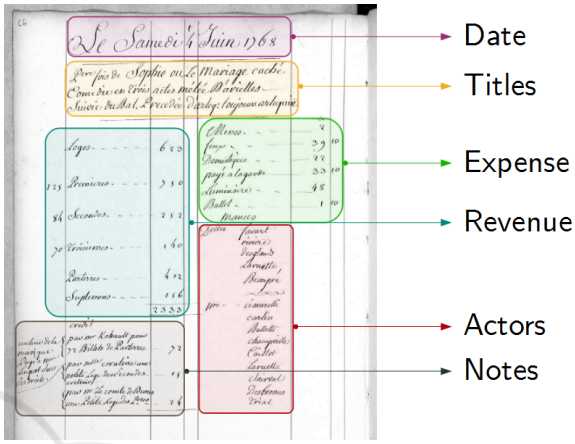


Figure 2: Example of a financial daily record for the Italian Comedy with identification fields.

4 CASE STUDY: RECORDS OF THE ITALIAN COMEDY

This paper deals with transcription and information extraction issues from historical documents with few or no annotated data. Our research aims at providing a handwriting recognition solution for documents of the Italian Comedy, from the 18^{th} century. They are provided by the BnF[1] (Bibliothèque nationale de France) as part of the ANR project CIRESFI. This data consists of more than 28,000 pages of financial records covering one century. Several evolutions were noticed within this dataset. The first one is related to the language which switches from Italian (with several dialects) to French. The second one is a change in the structure preserving the quantity of information. For one day, we can find the date, titles of the plays, revenues, expenses, actor names, and also some notes (as shown on Figure 2) but this layout fluctuates over the decades. Further works are in progress on the detection and segmentation of these fields for each page.

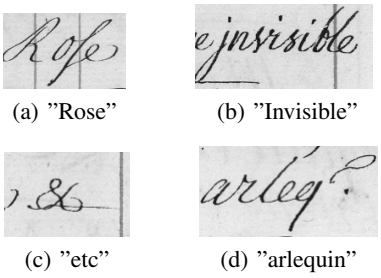
In this part of the project, our study focuses on the title field. This can be explained by the large collection of play titles of the Italian Comedy. The title field contains the list of plays that have been performed that day. Sometimes, this list gives more information, such as if it was a premiere, if it was played in special places, or in front of the king's court. An example of such complementary information is shown in Figure 2. The title field explains that this was the first performance of "Sophie ou le mariage caché" ("Sophie or the secret marriage") which was a comedy in three acts, preceded by "Arlequin toujours Arlequin" ("Arlequin always Arlequin"). Sometimes, actor names replace the names of their characters in the title. Thus, our collection of titles can not be considered as a ground truth but as a source of information.

The writing style is also an issue. At the beginning of the century, Italian actors wrote the records themselves. From the mid-century, there was only one writer for thirty years. Furthermore, there are differences as compared to the contemporary writing: special characters, as the long form of 's' (Figure 3(a)); evolution of the spelling, like using 'i' or 'j' indistinctly (Figure 3(b)); abbreviations such as "&c." for " etc " or symbols such as '○' (Figures 3(c) and 3(d)).

Thanks to a participative annotation website[2], we were able to collect information on the position of

[1] http://gallica.bnf.fr/accueil/

[2] http://recital.univ-nantes.fr/



(a) "Rose" (b) "Invisible"

(c) "etc" (d) "arlequin"

Figure 3: Special characters and abbreviations in the Italian Comedy documents.

the fields in the pages as well as their transcription. The algorithm of Seam Carving proposed by (Arvanitopoulos and Süsstrunk, 2014) has been applied to segment the blocks into title lines. Finally, we have manually validated the collected transcriptions of each block, each line, as well as the segmented lines themselves. Thus, we collected 971 validated lines with their transcriptions. To create the datasets for training, validation and test, we paid attention to distribute the lines of titles in both French and Italian among the various datasets. Moreover, title lines coming from the same page were not separated.

The aim of our work is not to provide another HWR method. Thus, we use CNN-BLSTM neural networks with CTC to produce knowledge transfer at two levels: from modern data to historical documents based on (Frinken et al., 2010) and from at least one language to another one (or more). We want to solve two problems: the first one is related to the language which is mainly in French; the second one is the old style of our data. This is why we use two types of data sources: the first one is data in French and the other one is data from the same period.

5 EXPERIMENTAL SETUP

5.1 Datasets

To apply transductive transfer learning, we need to carefully choose the data used to train our systems. To our knowledge, there are no available annotated data meeting all our criteria: from the 18^{th} century, in French and Italian, with a closed vocabulary on the Italian Comedy. Thus, each selected dataset has at least one characteristics in common with our data. The descriptions of the datasets are shown in Table 1.

Georges Washington (GW). The GW dataset (Fischer et al., 2012) contains 20 pages of letters from George Washington to his associates during the 18^{th} century. The writing style is very similar from letter to letter. The images are binarized and normalized. To be comparable to the state-of-the-art, the original partition of the dataset in train, validation and test is used in our work. In fact, as four partitions of the data exist, we randomly selected one of these partitions.

Los Esposalles (ESP). The ESP dataset (Romero et al., 2013) consists of 173 pages of old Spanish marriage records. The whole pages are provided with the segmentations and transcriptions of their words. There is only one writer for these pages. The raw images are used with a normalization of the height of 120 pixels (like the GW images).

RIMES (RM). The RM dataset (Augustin et al., 2006) is a French database used in several ICDAR competitions. It is composed of 12,723 pages of administrative letters written by 1,300 volunteers. The gray scale images are used with the same height normalization as the others. The official split for the ICDAR 2011 competition (Grosicki and El-Abed, 2011) provides 12,111 lines (including 11,333 lines for the training set) and 66,979 words (including 51,739 words for the training set). We keep this distribution.

Italian Comedy (CI). The CI dataset is a French and Italian dataset describing play titles. It is composed of 151 play titles. The images were also standardized at 120 pixels high and in gray scale too.

From all datasets, we have removed characters in the ground truth such as '#', '/' or '$' and replaced them by a "joker" character because they could not appear in the Italian Comedy. The accentuated characters and also the tied letter in the RIMES transcriptions are replaced by their simple form. Thus, we consider that the 'é' character is one form of 'e' like 's' has a long form and a short form in the 18^{th} century.

5.2 System Training

We have already detailed the architecture of our system in section 3 as well as some parameters. We carried out our experiments in successive stages. The first experiment allowed us to optimize the architecture of our system by evaluating it on two labeled datasets. With the best parameters obtained, we realized another experiment in order to select the best pairing of the dataset and to test the word recognition task on the Italian Comedy data. These results will be used for the third and final experiment which involves directly testing the transfer learning and the fine-tuning on our Italian Comedy data.

In transfer learning, the generalization capability of the classifier must be maximized: the classical technique of *early stopping* is used in order to select the best network. To avoid overfitting, training

Table 1: The selected datasets. In the upper part, each common point with the Italian Comedy is shown in bold. In the lower part, the data distribution on the training, validation and test sets is given as well as the name of the associated dataset.

Dataset		Georges Washington	Los Esposalles	RIMES	Italian Comedy
Language		English	Spanish	**French**	**French** (and Italian)
Period		**18th century**	**18th century**	21st century	**18th century**
Pixel value		Binarized	**Grayscale**	**Grayscale**	**Grayscale**
Words	Train	2,402 (GW_W)	45,102 (ESP_W)	51,739 (RM_W)	-
	Validation	1,199	5,637	7,464	-
	Test	1,292	5,637	7,776	-
Lines	Train	325 (GW_L)	-	11,333 (RM_L)	582 (CI_L)
	Validation	168	-	1,332	195
	Test	163	-	778	194

continues until the Negative Log-Likelihood (NLL) computed by the CTC is no longer decreasing during 20 epochs on all the validation datasets. One set of weights of the neural network is backed up for each validation dataset only if the NLL drops. The training is realized through all the parts of the network: feature extraction with the FCN, handwriting recognition with the BLSTM, and data labeling with the CTC.

Our first experiments on GW demonstrated that the training is performed more efficiently on the line images when it is gradually done. Hence, all images in the training set are sorted in an ascending order according to their label length. In this way, the training step goes from isolated characters to words, and finally to long lines. First, experiments exclusively run on the word datasets. Then, we extend them to the lines. This allows an increase in the performance of the BLSTM-CTC across all datasets.

To evaluate the performance of our system, we used the recognition rate at the character level (*CRR*) and at the word level (*WRR*). The *CRR* is defined by

$$CRR = \frac{N-(S+D+I)}{N},$$

with N the number of characters in the reference images, S the number of character substitutions, D the number of character deletions, and I the number of character insertions. The *WRR* is computed similarly on words. These two measures are case sensitive and treat the space as a character into a line. For the decoding step, no dictionary or language models were used on the three labeled datasets. Through our experiments we aims at defining a simple system capable of performing the transfer learning.

6 EXPERIMENTAL RESULTS

Previous experiments showed that the training phase is more efficient when it is gradually realized. In the learning from scratch on the lines of GW, after 100 epochs, the cost stays high and the recognition rate is about 10.7%. Nonetheless, if the network is fine tuned with the weights from the previous training on the word images, the recognition rate reaches 77.3% in 33 epochs. So, this gradual training has been used for the following experiments.

6.1 Optimizing the System Architecture

This part presents the optimization of the system architecture. Indeed, the parameters of each separate part of our system (CNN and BLSTM-CTC) have an impact on the quality of the transcriptions, as well as on the transfer phase. Those experiments aim at evaluating the various parameters that can influence our system, *i.e.* the CNN size in terms of number of extracted features and number of cells in the hidden LSTM layers.

The GW_W dataset is used more often for word spotting (Rath and Manmatha, 2007; Fischer et al., 2012; Frinken et al., 2012) than for HWR or word recognition (Lavrenko et al., 2004). The latter uses a HMM with a cross validation and 19 pages to train the system. It achieves a CRR of 56.8% without the out-of-vocabulary words. With only 15 pages to train our system, we achieve a CRR of 28.7%. In order to improve our results and deal with the limited amount of data in GW_W, we have paired it with the more extensive dataset RM_W, for each experiment.

Table 2 presents the results obtained by four systems trained on the $RM_W \cup GW_W$ training set and tested on RM_W and GW_W test sets. The systems differ from the number of features (32 or 128) and from the number of cells in their two LSTM layers (50 or 100). When the number of extracted features is set to 32, we can see that increasing the number of cells leads to a better rate on the characters. When the number of features is increased to 128, the character recogni-

Table 2: Results for several systems with the same training set, $RM_W \cup GW_W$.

Features	Cells	Test	CRR	WRR
32	50	GW_W	40.5	22.2
		RM_W	63.5	34.8
32	100	GW_W	43.9	22.2
		RM_W	67.6	39.9
128	100	GW_W	48.5	25.5
		RM_W	70.1	41.7

tion rate on GW_W evolves from 40.5% to 48.5%. The same observation can be made on RM_W, where the character recognition rate increases by 6.6%. However, the increase in the word recognition rate is not as important on the two test sets. Finally, these results allow us to conclude that the best configuration is obtained with 128 feature outputs from the CNN and with 100 LSTM cells, which was originally defined by (Graves and Schmidhuber, 2009). These settings will be kept for the rest of our experiments. We can also conclude that the addition of a large dataset helps to improve the results.

6.2 Optimizing the Word Datasets

Now that the architecture of the system has been set up, we have to define the datasets that will be used for training and testing our system. Among the three datasets, two must be dedicated to training and the third one must play the role of our CI data since we only have a limited amount of CI data with a ground-truth. The RIMES dataset must be part of the training dataset because it is the only one French dataset that we have here.

Table 3: Results obtain on word images for systems with 128 features and 100 cells for the LSTM layers.

Id	Train	Test	CRR	WRR
E_1	$RM_W \cup GW_W$	GW_W	48.5	25.5
		RM_W	70.1	41.7
		ESP_W	9.0	0.3
E_2	$RM_W \cup ESP_W$	GW_W	6.3	0.3
		RM_W	71.1	42.0
		ESP_W	91.1	75.9

Among studies on the RIMES dataset, (Pham et al., 2014) use a deep recurrent neural network composed of MDLSTM and CNN. The authors vary the number of LSTM cells from 30 to 200. They obtain a 84.9% character recognition rate with 50 cells, and a 84.2% character recognition rate with 100 cells. Progressively, our system (presented in Table 3) tends to reach this state-of-the-art. The high recognition rate obtained with RM and ESP on the words push us to select them for the last series of experiments on CI. These first experiments on transfer learning at the word level shows that it is a difficult task.

6.3 Experimenting Transfer Learning

In the previous experiments, we were in a phase of training and testing on labeled data. The last experiments deal with the learning transfer process on the line images, especially those of CI. It is interesting to evaluate the impact that can have the addition or not of target data during the training. Table 4 presents the result of three different experiments which include tests on CI. Each of them uses the saved weights from an experiment (indicated by an ID in the *train* column).

Table 4: Results for systems with 128 features, 100 cells for the LSTM layers and fine-tuned with the saved weights from the different experiences.

Id	Train	Test	CRR
E_3	$(E_2) \cup RM_L \cup ESP_W$	CI_L	10.6
E_4	$(E_2) \cup CI_L$	CI_L	25.5
E_5	$(E_3) \cup CI_L$	CI_L	28.7

Our user case corresponds to the first line. With the help of annotated data, we could achieve the results obtained on lines 2 and 3. At first, we add RM_L while keeping ESP_w and using the saved weights from the last experiment E_2 on the words presented in Table 3. In the experiments, we found out that fine-tuning is only useful when the resources used to set up the weights are still present because the system forgets. Then, these saved weights are used to perform fine-tuning on CI_L. This specialization of the network on the target data during the learning allows to increase the recognition rate of the characters by 15%. Finally, in order to observe the impact of the " space " character during the learning, the weights saved when learning on E_3 are chosen to perform fine-tuning on CI. It shows a 3.2% increase of the character recognition rate with respect to the rate obtained when the weights saved on E_2 were used.

Figures 4 and 5 show the curves obtained when fine-tuning the learning on CI_L, from $RM_W \cup ESP_W$ versus $RM_L \cup ESP_W$. In addition to improving the character recognition rate, we note that the learning is faster, 38 epochs against 27 epochs, when using RM_L instead of RM_W. Furthermore, the initial validation NLL is twice as low with RM_L and it quickly reaches a stability level, when the "space" character was learned before. With the fine-tuning on RM_W, the

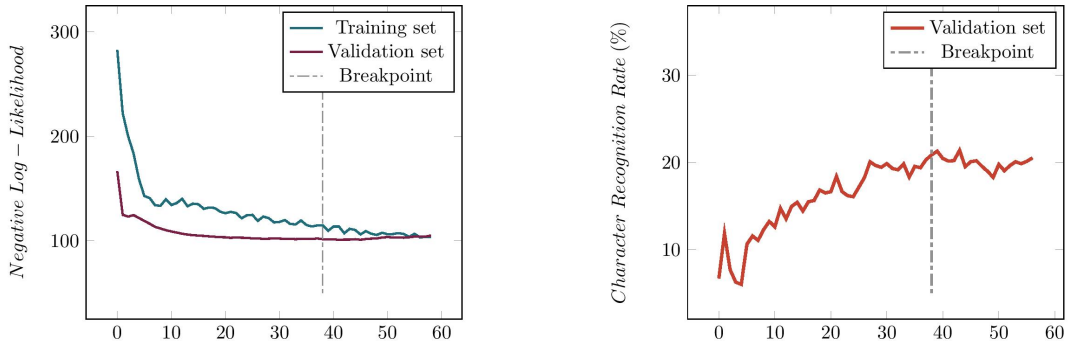


Figure 4: Fine-tuning of CI_L on E_2 : evolution of Log-Likelihood by *epoch*

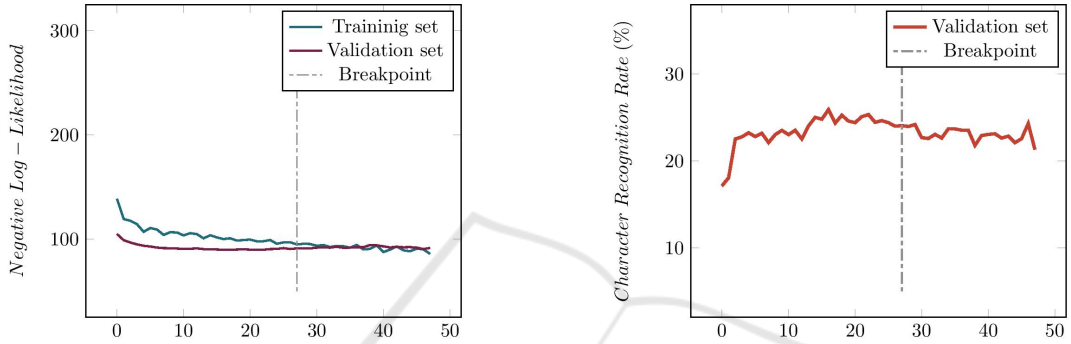


Figure 5: Fine-tuning of CI_L on E_3 evolution of Log-Likelihood by *epoch*

character recognition rate on the validation set grows gradually. We can notice a sharp fall at the beginning of the learning curve: it seems like the system forgets part of what it has learned to better specialize to CI_L. With RM_L, the character recognition rate curve behaves as the training continues as before the fine-tuning. It is also interesting to note that, here, the breakpoint on the validation NLL is just after the best value reached for the character recognition rate. Furthermore, we can see that, with just one iteration on a small set of target data, the character recognition rate already exceeds the first experiment we had conducted by testing directly on CI_L. Thus, a very simple system architecture without any language model helps to successfully achieve the learning transfer.

7 CONCLUSION

We performed experiments on *transductive transfer learning* for the task of handwriting recognition on a new historical dataset. Firstly, we optimized the parameters of the system to obtain the simplest and most performing system. Then, we defined the best pairing of datasets to realize the transfer learning. Moreover, it is necessary to pay attention to the balance of the representation of words in a dataset because this has a strong impact on the system. The last experiment with the CI data shows that even if both measures of recognition rate are low, we have a progression.

Our experiments allow us to conclude that HWR systems quickly specialize on learning data. We have found out that the addition of one or more resources makes it possible to improve character recognition. For now, it is still necessary to add a small amount of target data in the learning, to achieve a minimum recognition rate. In the long term, we want to build a system based on capitalization of resources in order to avoid those phases of annotation and manual validation. Our aim remains to be able to carry out transfer learning on data without prior costly knowledge.

These results will guide our future experiments. Besides the addition of resources and their balance control, we can explore several orientations. We will focus on solutions which do not require costly ground-truth. For example, using unsupervised training can take advantage of thousands unannotated available pages. This can be done by using an autoencoder for feature extraction layers instead of our FCN. Another low cost resource are dictionaries which can be collected from topic related corpus but this solution implies to add a language model post-processing of the network outputs. Finally we can also set up a deeper network with several LSTM layers as it is often used in the state-of-the-art. However, increasing the system size runs against the results presented

here to obtain a good recognition rate on data that are not used during learning. There is still some works to succeed in the recognition of new digitized documents from multilingual and multi-period resources.

REFERENCES

Arvanitopoulos, N. and Süsstrunk, S. (2014). Seam carving for text line extraction on color and grayscale historical manuscripts. In *ICFHR*, pages 726–731.

Augustin, E., Brodin, J.-m., Carré, M., Geoffrois, E., Grosicki, E., and Preteux, F. (2006). RIMES evaluation campaign for handwritten mail processing. In *ICFHR*.

Bunke, H., Roth, M., and Schukat-Talamazzini, E. G. (1995). Off-line cursive handwriting recognition using hidden markov models. *PR*, 28(9):1399–1413.

Cloppet, F., Eglin, V., Kieu, V., Stutzmann, D., and Vincent, N. (2016). ICFHR 2016 competition on the classification of medieval handwritings in latin script. In *ICFHR*, pages 590–595.

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, pages 2625–2634.

Fischer, A., Keller, A., Frinken, V., and Bunke, H. (2012). Lexicon-free handwritten word spotting using character HMMs. *PRL*, 33(7):934–942.

Fischer, A., Wüthrich, M., Liwicki, M., Frinken, V., Bunke, H., Viehhauser, G., and Stolz, M. (2009). Automatic transcription of handwritten medieval documents. In *VSMM*, pages 137–142.

Frinken, V., Fischer, A., Bunke, H., and Manmatha, R. (2010). Adapting BLSTM neural network based keyword spotting trained on modern data to historical documents. In *ICFHR*, pages 352–357.

Frinken, V., Fischer, A., Manmatha, R., and Bunke, H. (2012). A novel word spotting method based on recurrent neural networks. *IEEE Trans. on PAMI*, 34(2):211–224.

Graves, A. (2012). *Supervised Sequence Labelling with Recurrent Neural Networks*. Springer.

Graves, A. and Schmidhuber, J. (2009). Offline handwriting recognition with multidimensional recurrent neural networks. In *NIPS*, pages 545–552.

Grosicki, E. and El Abed, H. (2009). ICDAR 2009 handwriting recognition competition. In *ICDAR*, pages 1398–1402.

Grosicki, E. and El-Abed, H. (2011). ICDAR 2011: French handwriting recognition competition. In *ICDAR*, pages 1459–1463.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Koerich, A. L., Leydier, Y., Sabourin, R., and Suen, C. Y. (2002). A hybrid large vocabulary handwritten word recognition system using neural networks with hidden markov models. In *ICFHR*, pages 99–104.

Kozielski, M., Doetsch, P., Hamdani, M., and Ney, H. (2014). Multilingual off-line handwriting recognition in real-world images. In *DAS*, pages 121–125.

Lavrenko, V., Rath, T. M., and Manmatha, R. (2004). Holistic word recognition for handwritten historical documents. In *DIAL*, pages 278–287.

Lladós, J., Rusiñol, M., Fornés, A., Fernández, D., and Dutta, A. (2012). On the influence of word representations for handwritten word spotting in historical documents. *IJPRAI*, 26(05):1263002–1–25.

Moysset, B., Bluche, T., Knibbe, M., Benzeghiba, M. F., Messina, R., Louradour, J., and Kermorvant, C. (2014). The A2iA multi-lingual text recognition system at the second maurdor evaluation. In *ICFHR*, pages 297–302.

Murdock, M., Reid, S., Hamilton, B., and Reese, J. (2015). ICDAR 2015 competition on text line detection in historical documents. In *ICDAR*, pages 1171–1175.

Oprean, C., Likforman-Sulem, L., Popescu, A., and Mokbel, C. (2013). Using the web to create dynamic dictionaries in handwritten out-of-vocabulary word recognition. In *ICDAR*, pages 989–993.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Tran. on KDE*, 22(10):1345–1359.

Pham, V., Bluche, T., Kermorvant, C., and Louradour, J. (2014). Dropout improves recurrent neural networks for handwriting recognition. In *ICFHR*, pages 285–290.

Puigcerver, J., Toselli, A. H., and Vidal, E. (2015). ICDAR 2015 competition on keyword spotting for handwritten documents. In *ICDAR*, pages 1176–1180.

Rath, T. M. and Manmatha, R. (2007). Word spotting for historical documents. *IJDAR*, 9(2):139–152.

Romero, V., Fornés, A., Serrano, N., Sánchez, J. A., Toselli, A. H., Frinken, V., Vidal, E., and Lladós, J. (2013). The ESPOSALLES database: An ancient marriage license corpus for off-line hwr. *PR*, 46(6):1658–1669.

Senior, A. W. and Robinson, A. J. (1998). An off-line cursive handwriting recognition system. *PAMI*, 20(3):309–321.

Suryani, D., Doetsch, P., and Ney, H. (2016). On the benefits of convolutional neural network combinations in offline handwriting recognition. In *ICFHR*, pages 193–198.

Terasawa, K. and Tanaka, Y. (2009). Slit style HOG feature for document image word spotting. In *ICDAR*, pages 116–120.

Voigtlaender, P., Doetsch, P., and Ney, H. (2016). Handwriting recognition with large multidimensional long short-term memory recurrent neural networks. In *ICFHR*, pages 2228–233.