A Comparison of Sequence-Trained Deep Neural Networks and Recurrent Neural Networks Optical Modeling for Handwriting Recognition

Théodore Bluche[1,2], Hermann Ney[2,3], and Christopher Kermorvant[1]

[1] A2iA SA, Paris, France
[2] LIMSI CNRS, Spoken Language Processing Group, Orsay, France
[3] RWTH Aachen University, Human Language Technology and Pattern Recognition, Aachen, Germany

Abstract. Long Short-Term Memory Recurrent Neural Networks are the current state-of-the-art in handwriting recognition. In speech recognition, Deep Multi-Layer Perceptrons (DeepMLPs) have become the standard acoustic model for Hidden Markov Models (HMMs). Although handwriting and speech recognition systems tend to include similar components and techniques, DeepMLPs are not used as optical model in unconstrained large vocabulary handwriting recognition. In this paper, we compare Bidirectional LSTM-RNNs with DeepMLPs for this task. We carried out experiments on two public databases of multi-line handwritten documents: Rimes and IAM. We show that the proposed hybrid systems yield performance comparable to the state-of-the-art, regardless of the type of features (hand-crafted or pixel values) and the neural network optical model (DeepMLP or RNN).

Keywords: Handwriting Recognition •Recurrent Neural Networks • Deep Neural Networks

1 Introduction

Handwriting recognition is the problem of transforming an image into the text it contains. Unlike Optical Character Recognition (OCR), segmenting each character is difficult, mainly due to the cursive nature of handwriting. One usually prefers to recognize whole words or lines of text, i.e. the sequence of characters, with HMMs or RNNs.

In HMMs, the characters are modeled as sequences of hidden states, associated with an emission probability model. Gaussian Mixture Models (GMMs) is the standard optical model in HMMs. However, in the last decade, emission probability models based on artificial neural networks have (re)gained considerable interest in the community, mainly due to the *deep learning* trend in computer vision and speech recognition. In this latter domain, major improvements have been observed with the introduction of deep neural networks.

A significant usage of neural network for handwriting recognition should also be noted. The MNIST database of handwritten digits received a lot of

attention in computer vision and in the application of deep learning techniques. Convolutional Neural Networks introduced by Le Cun et al. [20] have soon been applied to handwriting recognition problems, and were recently tested on public databases for handwritten word recognition, yielding state-of-the-art results [5].

The state-of-the-art performance on many public handwriting databases is achieved by RNNs. This type of neural network has the ability to use more context than HMMs and to model the whole sequence directly. The best published results on IAM [19], Rimes [26, 19] and OpenHaRT [26, 31], were achieved by systems involving an RNN component.

In this work, we compare different approaches to optical modeling in handwriting recognition systems. In particular, we studied different kinds of neural networks (DeepMLPs and RNNs), and features (hand-crafted and pixel values).

We report results on the publicly available IAM [22] and Rimes [1] databases. Major improvements have been recently reported on these tasks, mainly due to a better pre-processing of the images, and an open-vocabulary language model [19]. This work shows that similar Word Error Rates (WERs) can be achieved with different kinds of features (hand-crafted geometric and statistical features, and pixel values), and optical models (DeepMLPs and RNNs), and a rather standard pre-processing. We note that for DeepMLPs to be comparable in performance to RNNs, a sequence training criterion, such as state-level Minimum Bayes Risk (sMBR) [18] should be used.

This paper is divided as follows. Section 2 contains a brief litterature review. Section 3 describes our systems. Section 4 presents the experiments carried out and the results obtained. Conclusions are drawn in Section 5.

2 Relation to Prior Work

Recurent Neural Networks, with the Long Short-Term Memory (LSTM) units, are particularly good for handwriting recognition. State-of-the-art systems for many public databases include an RNN component. Kozielski et al. [19] trained a bidirectional LSTM-RNN (BLSTM-RNN) on sequences of feature vectors, and HMM state targets. They then extract features from hidden layer activations to train a standard GMM-HMM, and report the best known results on the IAM database. On the other hand, Graves et al. [14], and more recently [4, 26] trained Multi-Dimensional LSTM-RNNs (MDLSTM-RNNs), which operate directly on the raw image, with a Connectionist Temporal Classification (CTC) objective, which allows to train the network directly using the sequence of characters as targets. With the dropout technique, [26] report the best results on both Rimes and OpenHaRT databases.

Multi-layer Perceptrons with one hidden layer were used for optical modeling in hybrid systems by España-Bocquera et al. [11] and Dreuw et al. [10]. Deep Neural Networks (DeepMLPs), were applied to simple handwriting recognition tasks such as isolated character or digits recognition [7, 8]. More recently, they were used in combination with HMMs for keywords spotting in handwritten documents [30]. They enjoy considerable research attention since efficient

training methods have been proposed. They achieve excellent results in various computer vision tasks (e.g. object recognition), but also in speech recognition, where they replace efficiently the conventional GMMs in HMMs. Their architecture is simple (multi-layer perceptrons), and their depth seems to contribute to better modeling [25] and robustness [9]. It has been shown [32, 28] that optimizing training criteria over whole sequences (e.g. sMBR), including the language constraints, leads to improvements compared with a framewise criterion. Similar (global) training of handwriting recognition systens were already proposed in the 90s [21, 20]. In this work, we show that the framework of DeepMLP and sequence training used in speech recognition can successfully be applied to handwriting recognition, with very good results on public databases, and compete with RNNs.

3 System Overview

3.1 Image Pre-Processing and Feature Extraction

The goal of pre-processing is to remove the undesirable variabilities from images. First, the lines are deskewed [3] and deslanted [6]. Then, the darkest 5% of pixels are mapped to black and the lightest 70% are mapped to white, with a linear interpolation in between, to enhance the contrast. We added 20 columns of white pixels to the beginning and end of each line to account for empty context. Most systems require an image with fixed height. We first detect three regions in the image (ascenders, descenders and core region) [33], and scale these regions to three fixed heights.

We built baseline systems using the handcrafted features described in [2], which gave reasonable performance on several public databases [23, 2]. We extracted them with a sliding window, scanned left-to-right through the preprocessed text line image. It is defined by two parameters: its width and shift (controlling the overlap between consecutive windows). To fix these parameters, we trained GMM-HMMs using the handcrafted features and different widths and shifts of the sliding window and keep the parameters yielding the best performance on the validation set. The optimal values we found are a width of 3px, a shift of 3px for both databases.

We also carried out experiments on pixel features (for NNs only). They are extracted with a sliding window of width 45px and shift 3px, rescaled to 20x32px. The pixel values are normalized to lie in the interval $[0, 1]$ (1 corresponding to white), producing 640-dimensional feature vectors. No Principal Component Analysis or other decorrelation or dimensionality reduction algorithm were applied.

3.2 Hidden Markov Models

The topology of the HMM is left-to-right: two output transitions per state, one to itself and one to the next state. We tried different number of HMM states in character models (along with different sliding window parameters), and kept

the values yielding the best GMM-HMM results on the validation sets. We built 6-state models for IAM and 5-state models for Rimes. We added two 2-state silence HMMs to model optional empty context on the left and right of words.

3.3 Neural Networks

Multi Layer Perceptrons and Deep Neural Networks Multi Layer Perceptrons (MLPs) are networks organized in several layers, each one fully connected to the next. The input corresponds to an observation vector, optionally concatenated with a small amount of previous and next frames. The output is a prediction of the HMM states. Deep Neural Networks (DeepMLPs) are MLPs with several hidden layers. We first initialize the weights with unsupervised pretraining, consisting in stacking Restricted Boltzmann Machines, trained with contrastive divergence, as explained in [15]. Then, we perform a supervised discriminative training of the whole network. The targets are obtained by forced alignment of the training set with a bootstrapping model. We optimize the cross-entropy criterion with Stochastic Gradient Descent (SGD).

Sequence training of neural networks consists in optimizing the network parameters with a sequence-discriminative criterion rather than using the frame-level cross-entropy criterion. Sequence training is similar to the discriminative training of GMM-HMMs. Among different possibilities, we chose the state-level Minimum Bayes Risk (sMBR) criterion, described in [18], which yields slightly better WER than other sequence criteria on a speech recognition task (Switchboard) [32]. In speech recognition, sequence training results in relative performance gains of 5-10% for various tasks [32, 29].

Recurrent Neural Networks (RNNs) In RNNs, the input to a given recurrent layer are not only the activations of the previous layers, but also its own activations at the previous time step. This characteristic enables them to naturally work with sequential inputs, and to use the past context to make predictions. Long Short-Term Memory (LSTM) units are recurrent neurons, in which a gating mechanism avoids the vanishing gradient problem, appearing in conventional RNNs [16, 14], and enables to learn arbitrarily long dependencies. In Bi-Directional LSTM-RNNs (BDLSTM-RNNs), LSTM layers are doubled: the second layer is connected to the "next" time step rather than the previous one. Thus the input sequence is processed in both directions, so past and future context are used to make predictions (see Fig. 1). The information coming from both directions is summed component-wise after the LSTM layers, and the result is an input for a feed-forward layer. This is a generalization of the MDLSTM-RNN architecture descibed in [14, 26] to the case of sequences of feature vectors.

Finally, the Connectionnist Temporal Classification (CTC) paradigm [13] has been used to train the RNNs. With CTC, no prior segmentation of the training data (line images) is required. Therefore, we do not need a bootstrapping procedure involving forced alignments with a previously trained HMM. Instead, we can select the target sequence to be the sequence of character in the image annotation, which simplifies the training procedure.

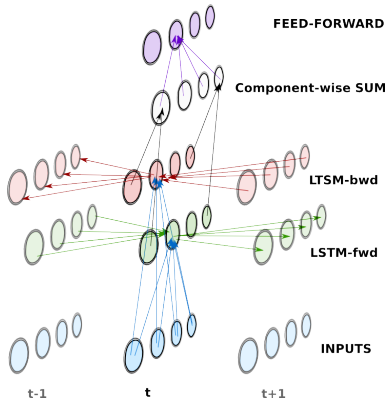

Fig. 1. Bidirectionnal Recurrent Neural Networks

4 Experiments and Results

4.1 Rimes and IAM Databases

The Rimes database [1] consists of images of handwritten paragraphs from simulated French mail. The setup for the ICDAR 2011 competition is a training set of 1,500 images, and an evaluation set of 100 images. We held out the last 149 images from the training set for system validation. We built a 4-gram language model (LM) with modified Kneser-Ney discounting from the training annotations. The vocabulary is made of 12k words. The language model has a perplexity of 18 and out-of-vocabulary (OOV) rate of 2.9% on the validation set (18 and 2.6% on the evaluation set).

The IAM database [22] consists of images of handwritten documents. They correspond to English texts exctracted from the LOB corpus [17], copied by different writers. The database is split into 747 images for training, 116 for validation, and 336 for evaluation. We used a 3-gram language model limited to the 50k most frequent words from the training set. It was trained on the LOB, Brown and Wellington corpora. The passages of the LOB corpus appearing in the validation and evaluation sets were removed prior to LM training. The resulting model has a perplexity of 298 and OOV rate of 4.3% on the validation set (329 and 3.7% on the evaluation set).

4.2 Decoding Method

We used the Kaldi toolkit [27] to decode the sequences of observation vectors (GMMs, DeepMLPs), or the sequences of character predictions (RNNs). The decoding was done for complete paragraphs rather than lines, to benefit from the language model history across line boundaries. The optical scaling factor, balancing the importance given to the optical model scores and to the language

model scores, and the word insertion penalty were tuned on the validation sets. This optimization can yield from 1 to 3% absolute improvement.

4.3 GMM-HMM

We trained GMM-HMM on both tasks, using the handcrafted features, and the Maximum Likelihood criterion. The number of Gaussians in the mixtures was increased at each iteration until the performance on the validation set decreases for more than 5 iterations. The GMM-HMM have not been discriminatively trained. They were only used to bootstrap the training of DeepMLPs.

4.4 Deep Neural Networks

To train the DeepMLPs, we performed the forced alignments of the training set with the GMM-HMMs, to have a target HMM state for each input observation. We held out 10% of this dataset for validation and early stopping. Overall, the datasets contain 5,6M examples for Rimes and 3,8M examples for IAM.

DeepMLP on Handcrafted Features We inverstigated different numbers of hidden layers (1 to 7) in the DeepMLP and different sizes of input context ($\pm\{1, 3, 5, 7, 9\}$ frames). The number of hidden nodes in each layer was set to 1,024. The input features were normalized to zero mean and unit variance along each dimension. The networks were pre-trained using 1 epoch of unsupervised training for each layer, followed by a few epochs of supervised training with stochastic gradient descent and a cross-entropy criterion. The training finished when no more improvement was observed on the validation set.

The results are depicted on Fig. 2. The performances of the different networks are similar to each other. It looks like more than one hidden layer is generally better, but the performance gain when we add more layers is not significant. We selected the best architectures based on the performance on the validation sets: 5 hidden layers with 1,024 units and 15 frames of context (central frame ± 7) for IAM, 4 hidden layers with 1,024 units and 7 frames of context (central frame ± 3) for Rimes. Additionally, training the networks with 5 more epochs of sMBR sequence training allowed to obtain 4 to 6% relative WER improvement (Table 1).

DeepMLP on Pixels Instead of adding context frames to the central frames, we extracted the pixels values in a larger sliding window. The means and standard deviations were computed across all dimensions simultaneously, not separately.

For the pixel DeepMLP, we notice a wider difference between one and more hidden layers (Fig. 3) than for DeepMLP on handcrafted features. The justification could be that in hancrafted features DeepMLPs, the inputs are already a higher level representation of the image, while in pixel DeepMLPs, the first

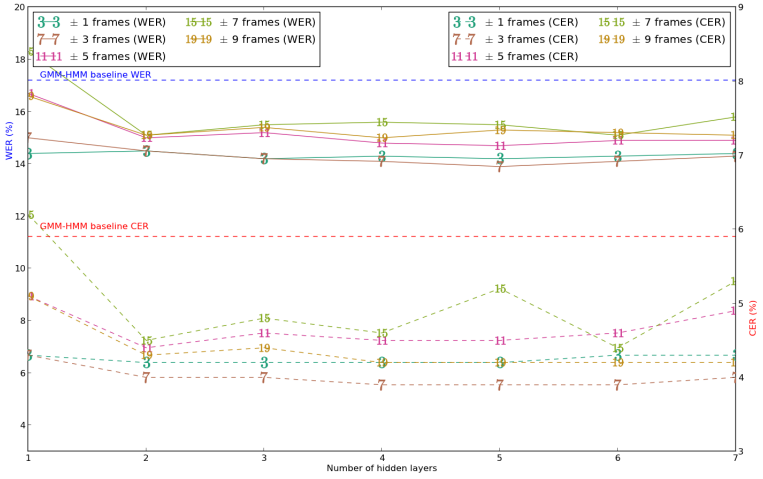

Fig. 2. Effect of depth and size of context (Features DeepMLP, Rimes validation set)

Table 1. Improvement brought by sMBR sequence training (results reported on validation sets)

System	WER - Rimes		CER - Rimes		WER - IAM		CER - IAM	
Features DeepMLP	14.1%		4.0%		12.4%		4.1%	
+ sMBR training	13.5%	(-4.2%)	4.0%	(-0.0%)	11.7%	(-5.6%)	3.9%	(-4.9%)
Pixel DeepMLP	13.6%		3.9%		12.4%		4.4%	
+ sMBR training	13.1%	(-3.7%)	3.8%	(-2.6%)	11.8%	(-4.8%)	4.2%	(-4.5%)

layer(s) perform the transformation of the image into a higher level representation. We selected the best architectures based on the performance on the validation sets: 4 hidden layers with 1,024 units for IAM, 7 hidden layers with 1,024 units for Rimes. Again, sMBR training brought a few percents relative improvement over cross-entropy training (Table 1).

4.5 Recurrent Neural Networks

Since the RNNs are trained with a CTC objective function to predict sequences of characters, there is no need for a bootstrapping procedure. All the RNNs have been trained on the whole training set and validated on the validation set.

BDLSTM-RNN on Handcrafted Features The RNNs naturally takes into account the left and right context to make predictions. Thus, we did not concatenate context feature frames. The input features were normalized to zero mean and unit variance along each dimension. We explored different depths and

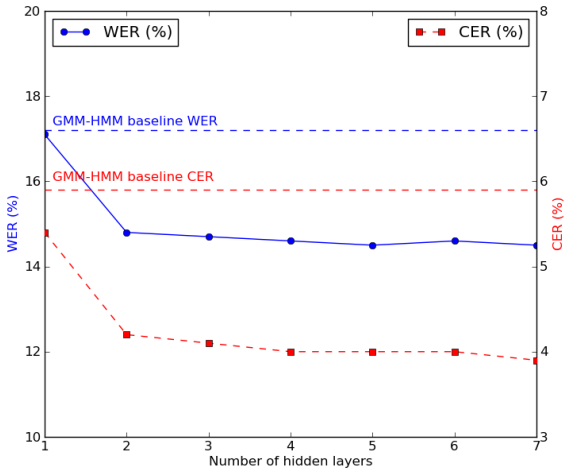

Fig. 3. Effect of increasing the number of hidden layers (Pixel DeepMLP, Rimes validation set)

widths, and also applied the dropout regularization technique in feed-forward layers, as explained in [26].

We report the results in Table 2. uCER stands for unconstrained CER, and refers to the character error rate when the RNN is used alone to make character predictions, i.e. without lexical and language model. While it seems better to have more than one hidden layer, the biggest improvements were achieved with dropout. The best architectures, selected based on the results on the validation sets, are 7 hidden layers (4 LSTM and 3 feed-forward) of 200 units with dropout for Rimes and IAM.

BDLSTM-RNN on Pixels For pixel features, the inputs are normalized with the mean and standard deviation of pixel values across all dimensions. We also explored different widths and depths and dropout, and selected the best models based on the validation results. For Rimes and IAM, the best network has 7 hidden layers of 200 units and dropout. The results for different architectures on Rimes database are shown on Table 2. Again, we notice that the effect of having more than one hidden layer is more important for pixel-based models than for models using handcrafted features.

The final results, comparing different models and input features on the one hand, and comparing our proposed systems with other published results on the other hand, are reported on Tables 3 (IAM) and 4 (Rimes). We see that both handcrafted and pixel features, and both DeepMLPs and RNNs can achieve results that are close to the best reported ones. For DeepMLPs, sequence training seems crucial to attain this performance. Furthermore, we notice that although RNNs have become a standard component of handwriting recognition systems,

Table 2. RNNs on handcrafted and pixel features, results for Rimes validation set. uCER stands for unconstrained CER. WER and CER are computed with a lexicon and language model.

		Handcrafted features				**Pixel features**			
Archi.	dropout	CTC cost	uCER	WER	CER	CTC cost	uCER	WER	CER
1x100	-	0.5217	14.5	14.9	4.7	1.201	33.8	24.1	10.3
3x100	-	0.3864	10.6	13.6	4.1	0.4834	12.9	15.1	5.1
5x100	-	0.3516	9.3	14.7	4.3	0.3637	9.8	14.0	4.4
5x200	-	0.3295	8.5	13.5	3.9	0.3724	9.7	15.4	4.9
7x100	-	0.3093	8.0	13.8	4.1	0.3313	8.7	14.5	4.5
7x200	-	0.2969	8.0	14.1	4.1	0.3445	8.9	14.7	5.0
7x200	x	0.2397	5.7	**12.7**	**3.6**	0.2351	6.0	**13.6**	**4.1**
9x100	-	0.2937	7.6	13.2	3.9	0.3229	8.6	14.5	4.5
9x100	x	0.2565	6.0	13.1	3.8	0.2559	6.3	13.8	4.4

Table 3. Results on IAM database

		Dev.		**Eval.**	
		WER	**CER**	**WER**	**CER**
g.	GMM-HMM baseline	15.2	6.3	19.6	9.0
df.	Features DeepMLP-5x1024	11.7	3.9	14.7	5.8
dp.	Pixel DeepMLP-4x1024	11.8	4.2	14.7	5.9
rf.	Feature BDLSTM-RNN 7x200 + dropout	11.9	3.9	14.3	5.3
rp.	Pixels BDLSTM-RNN 7x200 + dropout	11.8	4.0	14.8	5.6
	ROVER rf + rp + df + dp	9.7	3.6	**11.9**	**4.9**
	Kozielski et al. [19]	**9.5**	**2.7**	13.3	5.1
	Pham et al. [26]	11.2	3.7	13.6	5.1
	Kozielski et al. [19]	11.9	3.2	-	-

DeepMLPs – which have become standard in hybrid speech recognition systems – can perform equally well. Finally, we cannot draw a clear conclusion regarding whether RNNs or DeepMLPs should be preferred, or whether handcrafted features are more suited than pixel values.

Our different optical models and features are also complementary, as shown by their ROVER combination [12], which, to the best of our knowledge constitute the best published results on both databases, outperforming the open-vocabulary approaches proposed in [19] and [24].

5 Conclusion

In this paper, we shown that state-of-the-art WERs can be achieved with both DeepMLPs - standard method for speech recognition, and RNNs - standard

Table 4. Results on Rimes database

		Dev.		Eval.	
		WER	**CER**	**WER**	**CER**
g.	GMM-HMM baseline	17.2	5.9	15.8	6.0
df.	Features DeepMLP-4x1024	13.5	4.0	13.5	4.1
dp.	Pixel DeepMLP-7x1024	13.1	3.8	12.9	3.8
rf.	Feature BDLSTM-RNN 7x200 + dropout	12.7	3.6	12.7	4.0
rp.	Pixels BDLSTM-RNN 7x200 + dropout	13.6	4.1	13.8	4.3
	ROVER rf + rp + df + dp	11.8	3.4	**11.8**	3.7
	Pham et al. [26]	-	-	12.3	**3.3**
	Messina et al. [24]	-	-	13.3	-
	Kozielski et al. [19]	-	-	13.7	4.6

method for handwriting recognition. Even with a pretty simple image preprocessing, the pixel values could replace handcrafted features. Future work may include an evaluation of convolutional neural networks and Multi-Dimensional (MD)LSTM-RNNs for a more comprehensive comparison of neural network optical modeling. An evaluation of a tandem combination (where the neural networks are used to extract features rather than to make predictions) could be carried out. Finally, it would be interesting to evaluate the robustness of the proposed models, i.e. to see how good the results could be when these systems are applied to new databases, not seen during training.

References

1. Augustin, E., Carré, M., Grosicki, E., Brodin, J.M., Geoffrois, E., Preteux, F.: RIMES evaluation campaign for handwritten mail processing. In: Proceedings of the Workshop on Frontiers in Handwriting Recognition (2006)
2. Bianne, A.L., Menasri, F., Al-Hajj, R., Mokbel, C., Kermorvant, C., Likforman-Sulem, L.: Dynamic and Contextual Information in HMM modeling for Handwriting Recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence 33(10), 2066 – 2080 (2011)
3. Bloomberg, D.S., Kopec, G.E., Dasari, L.: Measuring document image skew and orientation. In: IS&T/SPIE's Symposium on Electronic Imaging: Science & Technology. pp. 302–316. International Society for Optics and Photonics (1995)
4. Bluche, T., Louradour, J., Knibbe, M., Moysset, B., Benzeghiba, M., Kermorvant, C.: The A2iA Arabic Handwritten Text Recognition System at the OpenHaRT2013 Evaluation. In: 11th IAPR Workshop on Document Analysis Systems (DAS2014). pp. 161–165 (2014)
5. Bluche, T., Ney, H., Kermorvant, C.: Tandem HMM with convolutional neural network for handwritten word recognition. In: 38th International Conference on Acoustics Speech and Signal Processing (ICASSP2013). pp. 2390 – 2394 (2013)
6. Buse, R., Liu, Z.Q., Caelli, T.: A structural and relational approach to handwritten word recognition. IEEE Transactions on Systems, Man and Cybernetics 27(5), 847–61 (Jan 1997)

7. Ciresan, D.C., Meier, U., Gambardella, L.M., Schmidhuber, J.: Deep, big, simple neural nets for handwritten digit recognition. Neural computation 22(12), 3207–3220 (2010)
8. Cireşan, D.C., Meier, U., Gambardella, L.M., Schmidhuber, J.: Deep big multilayer perceptrons for digit recognition. In: Neural Networks: Tricks of the Trade, pp. 581–598. Springer (2012)
9. Deng, L., Li, J., Huang, J.T., Yao, K., Yu, D., Seide, F., Seltzer, M., Zweig, G., He, X., Williams, J., et al.: Recent advances in deep learning for speech research at microsoft. In: 38th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2013). pp. 8604–8608. IEEE (2013)
10. Dreuw, P., Doetsch, P., Plahl, C., Ney, H.: Hierarchical hybrid mlp/hmm or rather mlp features for a discriminatively trained gaussian hmm: a comparison for offline handwriting recognition. In: 18th IEEE International Conference on Image Processing (ICIP2011). pp. 3541–3544. IEEE (2011)
11. Espana-Boquera, S., Castro-Bleda, M.J., Gorbe-Moya, J., Zamora-Martinez, F.: Improving offline handwritten text recognition with hybrid hmm/ann models. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(4), 767–779 (2011)
12. Fiscus, J.G.: A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover). In: IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU1997). pp. 347–354. IEEE (1997)
13. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd international conference on Machine learning. pp. 369–376. ACM (2006)
14. Graves, A., Schmidhuber, J.: Offline handwriting recognition with multidimensional recurrent neural networks. In: NIPS. pp. 545–552 (2008)
15. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural computation 18(7), 1527–1554 (2006)
16. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9(8), 1735–1780 (1997)
17. Johansson, S.: The lob corpus of british english texts: presentation and comments. ALLC journal 1(1), 25–36 (1980)
18. Kingsbury, B.: Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009). pp. 3761–3764. IEEE (2009)
19. Kozielski, M., Doetsch, P., Ney, H.: Improvements in RWTH's system for off-line handwriting recognition. In: International Conference on Document Analysis and Recognition (ICDAR2013). pp. 935 – 939 (2013)
20. Le Cun, Y., Bottou, L., Bengio, Y.: Reading checks with multilayer graph transformer networks. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP1997). vol. 1, pp. 151–154. IEEE (1997)
21. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11), 2278–2324 (1998)
22. Marti, U.V., Bunke, H.: The iam-database: an english sentence database for offline handwriting recognition. International Journal on Document Analysis and Recognition 5(1), 39–46 (2002)
23. Menasri, F., Louradour, J., Bianne-Bernard, A.L., Kermorvant, C.: The a2ia french handwriting recognition system at the rimes-icdar2011 competition. In: IS&T/SPIE Electronic Imaging. pp. 82970–82970. International Society for Optics and Photonics (2012)

24. Messina, R., Kermorvant, C.: Surgenerative Finite State Transducer n-gram for Out-Of-Vocabulary Word Recognition. In: 11th IAPR Workshop on Document Analysis Systems (DAS2014). pp. 212–216 (2014)
25. Mohamed, A.r., Hinton, G., Penn, G.: Understanding how deep belief networks perform acoustic modelling. In: 37th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2012). pp. 4273–4276. IEEE (2012)
26. Pham, V., Bluche, T., Kermorvant, C., Louradour, J.: Dropout improves Reccurent Neural Networks for Handwriting Recognition. In: 14th International Conference on Frontiers in Handwriting Recognition (ICFHR2014) (2014)
27. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al.: The kaldi speech recognition toolkit. In: Workshop on Automatic Speech Recognition and Understanding (ASRU2011). pp. 1–4 (2011)
28. Sainath, T.N., Mohamed, A.r., Kingsbury, B., Ramabhadran, B.: Deep convolutional neural networks for lvcsr. In: 38th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2013). pp. 8614–8618. IEEE (2013)
29. Su, H., Li, G., Yu, D., Seide, F.: Error back propagation for sequence training of context-dependent deep networks for conversational speech transcription. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2013). pp. 6664–6668 (2013)
30. Thomas, S., Chatelain, C., Paquet, T., Heutte, L.: Un modèle neuro markovien profond pour l'extraction de séquences dans des documents manuscrits. Document numérique 16(2), 49–68 (2013)
31. Tong, A., Przybocki, M., Maergner, V., El Abed, H.: NIST 2013 Open Handwriting Recognition and Translation (OpenHaRT13) Evaluation. In: 11th IAPR Workshop on Document Analysis Systems (DAS2014). pp. 81–85 (2014)
32. Veselý, K., Ghoshal, A., Burget, L., Povey, D.: Sequence-discriminative training of deep neural networks. In: 14th Annual Conference of the International Speech Communication Association (INTERSPEECH2013). pp. 2345–2349 (2013)
33. Vinciarelli, A., Luettin, J.: A new normalisation technique for cursive handwritten words. Pattern Recognition Letters 22, 1043–1050 (2001)

Acknowledgments

The authors would like to thank Michal Kozielsky and his colleagues from RWTH for providing the language model used in IAM experiments. This work was partly achieved as part of the Quaero Program, funded by OSEO, French State agency for innovation and was supported by the French Research Agency under the contract Cognilego ANR 2010-CORD-013.