

# AttentionHTR: Handwritten Text Recognition Based on Attention Encoder-Decoder Networks

Dmitrijs Kass<sup>1</sup> and Ekta Vats<sup>2</sup>[0000–0003–4480–3158]

<sup>1</sup> Department of Information Technology, Uppsala University, Sweden  
dmitrijs.kass@it.uu.se

<sup>2</sup> Centre for Digital Humanities, Department of ALM, Uppsala University, Sweden  
ekta.vats@abm.uu.se

**Abstract.** This work proposes an attention-based sequence-to-sequence model for handwritten word recognition and explores transfer learning for data-efficient training of HTR systems. To overcome training data scarcity, this work leverages models pre-trained on scene text images as a starting point towards tailoring the handwriting recognition models. ResNet feature extraction and bidirectional LSTM-based sequence modeling stages together form an encoder. The prediction stage consists of a decoder and a content-based attention mechanism. The effectiveness of the proposed end-to-end HTR system has been empirically evaluated on a novel multi-writer dataset Imgur5K and the IAM dataset. The experimental results evaluate the performance of the HTR framework, further supported by an in-depth analysis of the error cases. Source code and pre-trained models are available at GitHub<sup>3</sup>.

**Keywords:** Handwritten Text Recognition · attention encoder-decoder networks · sequence-to-sequence model · transfer learning · multi-writer.

## 1 Introduction

Historical archives and cultural institutions contain rich heritage collections from historical times that are to be digitized to prevent degradation over time. The importance of digitization has led to a strong research interest in designing methods for automatic handwritten text recognition (HTR). However, handwritten text possesses variability in handwriting styles, and the documents are often heavily degraded. Such issues render the digitization of handwritten material more challenging and suggest the need to have more sophisticated HTR systems.

The current state-of-the-art in HTR is dominated by deep learning-based methods that require a significant amount of training data. Further, to accurately model the variability in writing styles in a multi-writer scenario, it is important to train the neural network on a variety of handwritten texts. However, only a limited amount of annotated data is available to train an end-to-end HTR model from scratch, and that affects the performance of the HTR system. This

<sup>3</sup> <https://github.com/dmitrijsk/AttentionHTR>

work proposes an end-to-end HTR system based on attention encoder-decoder architecture and presents a transfer learning-based approach for data-efficient training. In an effort toward designing a data-driven HTR pipeline, a Scene Text Recognition (STR) benchmark model [1] is studied, which is trained on nearly 14 million synthetic scene text word images. The idea is to fine-tune the STR benchmark model on different handwritten word datasets.

The novelty and technical contributions of this work are as follows. (a) The goal of this work is to leverage transfer learning to enable HTR in cases where training data is scarce. (b) Scene text images and a new dataset, Imgur5K [13], are used for transfer learning. The Imgur5K dataset contains a handwritten text by approximately 5000 writers, which allows our proposed model to generalize better on unseen examples. (c) Our transfer learning-based framework produces a model that is applicable in the real world as it was trained on word examples from thousands of authors, with varying imaging conditions. (d) The proposed attention-based architecture is simple, modular, and reproducible, more data can be easily added in the pipeline, further strengthening the model’s accuracy. (e) A comprehensive error analysis using different techniques such as bias-variance analysis, character-level analysis, and visual analysis is presented. (f) An ablation study is conducted to demonstrate the importance of the proposed framework.

## 2 Related work

In document analysis literature, popular approaches towards handwriting recognition include Hidden Markov Models [17], Recurrent Neural Networks (RNN) [7], CNN-RNN hybrid architectures [6], and attention-based sequence to sequence (seq2seq) models [3,12,20]. Recent developments [11,15] suggest that the current state-of-the-art in HTR is advancing towards using attention-based seq2seq models. RNNs are commonly used to model the temporal nature of the text, and the attention mechanism performs well when used with RNNs to focus on the useful features at each time step [12].

Seq2seq models follow an encoder-decoder architecture, where one network encodes an input sequence into a fixed-length vector representation, and the other decodes it into a target sequence. However, using a fixed-length vector affects the performance, and therefore attention mechanisms are gaining importance due to enabling an automatic search for the most relevant elements of an input sequence to predict a target sequence [2].

This work investigates attention-based encoder-decoder networks [1] for handwriting recognition, based on ResNet for feature extraction, bidirectional LSTMs for sequence modeling, a content-based attention mechanism for prediction [2], and transfer learning from STR domain to HTR to overcome lack of training data. To this end, the STR benchmark [1] has been studied, and the pre-trained models from STR are used in this work to fine-tune the HTR models. In [1], the models were trained end-to-end on the union of MJSynth [10] and SynthText [8] datasets. After filtering out words containing non-alphanumeric characters, these datasets contain 8.9M and 5.5M word-level images for training, respectively.

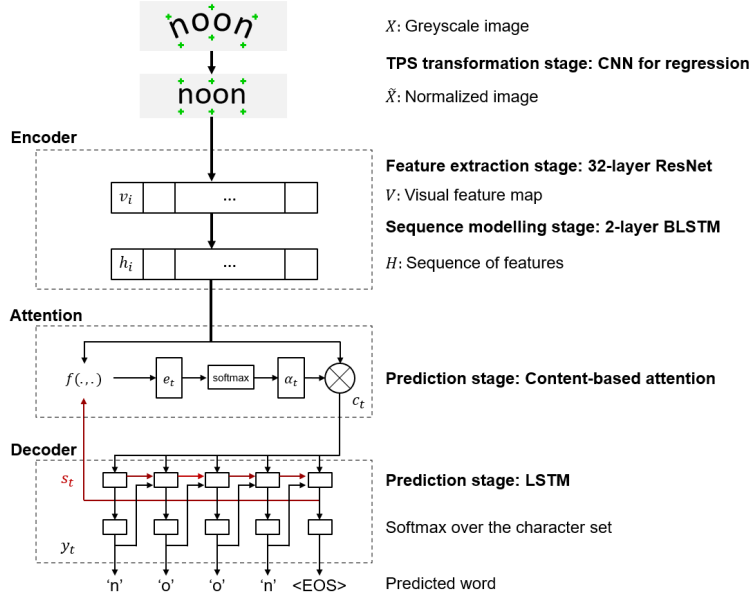
The related methods include [3,11,12,15,20], where [3] is the first attempt towards using an attention-based model for HTR. The limitation of this work is that it requires pre-training of the features extracted from the encoding stage using the Connectionist Temporal Classification (CTC) loss to be relevant. The method proposed in [20] is based on a sliding window approach. Limitations of this work include an increase in overhead due to sliding window size initialization and parameter tuning. The model proposed in [12] consists of an encoder, an attention mechanism, and a decoder. The task of the encoder is to extract deep visual features and it is implemented as a combination of two neural networks: a Convolutional Neural Network (CNN) and a bidirectional Gated Recurrent Unit (BGRU). The task of the decoder is to predict one character at a time. Each time a character is being predicted an attention mechanism automatically defines a context as a combination of input features that are the most relevant for the current predicted character and improves the predictive performance of the decoder. Similar to the proposed method, [12] does not require pre-processing, a pre-defined lexicon, or a language model. The method proposed in [15] combines CNN and a deep bidirectional LSTM, and also investigates various attention mechanisms and positional encodings to address the problem of input-output text sequence alignment. For decoding the actual character sequence, it uses a separate RNN. Recently, [11] introduced Candidate Fusion as a novel approach toward integrating a language model into a seq2seq architecture. The proposed method performs well in comparison with the related methods using the IAM dataset under the same experimental settings.

### 3 Attention-based encoder-decoder network

The overall pipeline of the proposed HTR method is presented in Fig. 1. The model architecture consists of four stages: transformation, feature extraction, sequence modeling, and prediction, discussed as follows.

**Transformation stage.** Since the handwritten words appear in irregular shapes (e.g. tilted, skewed, curved), the input word images are normalized using the thin-plate spline (TPS) transformation [4]. The localization network of TPS takes an input image and learns the coordinates of fiducial points that are used to capture the shape of the text. Coordinates of fiducial points are regressed by a CNN and the number of points is a hyper-parameter. TPS also consists of a grid generator that maps the fiducial points on the input image  $\mathbf{X}$  to the normalized image  $\tilde{\mathbf{X}}$ , and then a sampler interpolates pixels from the input to the normalized image.

**Feature extraction stage.** A 32-layer residual neural network (ResNet) [9] is used to encode a normalized  $100 \times 32$  grayscale input image into a 2D visual feature map  $\mathbf{V} = \{\mathbf{v}_i\}, i = 1..I$ , where  $I$  is the number of columns in the feature map. The output visual feature map has  $512 \text{ channels} \times 26 \text{ columns}$ . Each column of the feature map corresponds to a receptive field on the transformed input image. Columns in the feature map are in the same order as receptive fields on the image from left to right [18].



**Fig. 1.** Attention-based encoder-decoder architecture for HTR.

**Sequence modeling stage.** The features  $\mathbf{V}$  from the feature extraction stage are reshaped into sequences of features  $\mathbf{H}$ , where each column in a feature map  $\mathbf{v}_i \in \mathbf{V}$  is used as a sequence frame [18]. A bidirectional LSTM (BLSTM) is used for sequence modeling which allows contextual information within a sequence (from both sides) to be captured, and renders the recognition of character sequences simpler and more stable as compared to recognizing each character independently. For example, a contrast between the character heights in “il” can help recognize them correctly, as opposed to recognizing each character independently [18]. Finally, two BLSTMs are stacked to learn a higher level of abstractions. Dimensions of output feature sequence are  $256 \times 26$ .

**Prediction stage.** An attention-based decoder is used to improve character sequence predictions. The decoder is a unidirectional LSTM and attention is content-based. The prediction loop has  $T$  time steps, which is the maximum length of a predicted word. At each time step  $t = 1..T$  the decoder calculates a probability distribution over the character set

$$\mathbf{y}_t = \text{softmax}(\mathbf{W}_0 \mathbf{s}_t + \mathbf{b}_0)$$

where  $\mathbf{W}_0$  and  $\mathbf{b}_0$  are trainable parameters and  $\mathbf{s}_t$  is a hidden state of decoder LSTM at time  $t$ . A character with the highest probability is used as a prediction. The decoder can generate sequences of variable lengths, but the maximum length is fixed as a hyper-parameter. We used the maximum length of 26 characters, including an end-of-sequence (EOS) token that signals the decoder to stop making predictions after EOS is emitted. The character set is also fixed. A case-

insensitive model used further in experiments uses an alphanumeric character set of length 37, which consists of 26 lower-case Latin letters, 10 digits, and an EOS token. A character set of a case-sensitive model also includes 26 upper-case Latin letters and 32 special characters ( $\sim \wedge \backslash ! " \# \$ \% \& ' ( ) * + , - . / : ; < = > ? @ [ \ ] \{ \} \}$ ), and has a total length of 95 characters.

A hidden state of the decoder LSTM  $\mathbf{s}_t$  is conditioned on the previous prediction  $\mathbf{y}_{t-1}$ , a context vector  $\mathbf{c}_t$  and a previous hidden state

$$\mathbf{s}_t = LSTM(\mathbf{y}_{t-1}, \mathbf{c}_t, \mathbf{s}_{t-1}).$$

A context  $\mathbf{c}_t$  is calculated as a weighted sum of encoded feature vectors  $\mathbf{H} = \mathbf{h}_1, \dots, \mathbf{h}_I$  from the sequence modeling stage as

$$\mathbf{c}_t = \sum_{i=1}^I \alpha_{ti} \mathbf{h}_i,$$

where  $\alpha_{ti}$  are normalized attention scores, also called attention weights or alignment factors, and are calculated as

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^I \exp(e_{tk})},$$

where  $e_{ti}$  are attention scores calculated using Bahdanau alignment function [2]

$$e_{ti} = f(\mathbf{s}_{t-1}, \mathbf{h}_i) = \mathbf{v}^\top \tanh(\mathbf{W}\mathbf{s}_{t-1} + \mathbf{V}\mathbf{h}_i + \mathbf{b}),$$

where  $\mathbf{v}$ ,  $\mathbf{W}$ ,  $\mathbf{V}$  and  $\mathbf{b}$  are trainable parameters.

**Transfer learning.** Transfer learning is used in this work to resolve the problem of insufficient training data. In general, it relaxes the assumption that the training and test datasets must be identically distributed [21] and tries to transfer the learned features from the source domain to the target domain. The source domain here is STR, and the target domain is HTR. This paper uses the strategy of fine-tuning [21] by initializing the weights from a pre-trained model and continuing the learning process by updating all 49.6M parameters in it. The architecture of the fine-tuned model is the same as that of the pre-trained model. This, together with a significant overlap between the source and the target domains (synthetic scene text and handwritten text), allows the use of training datasets that are relatively small in size (up to 213K), compared to the training dataset used to obtain the pre-trained model (14.4M).

The learning rate is an important parameter to consider in fine-tuning, and its impact on the results is examined. A high learning rate can make the model perform poorly by completely changing the parameters of a pre-trained model. And a low learning rate can adjust a pre-trained model's parameters to a new dataset without making many adjustments to the original weights. A batch size of 32 has been selected based on the size of the training set and the memory requirements. Early stopping is used as a regularization method to avoid over-fitting due to its effectiveness and simplicity.

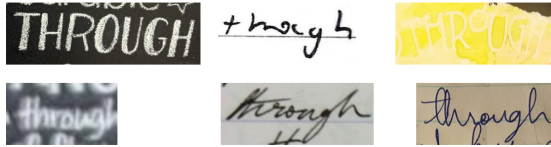


Fig. 2. Imgur5K dataset samples. Top row: training set. Bottom row: test set.

## 4 Experimental results

### 4.1 Datasets

Two word-level multi-writer datasets, IAM [14] and Imgur5K [13], are used. Raw image files were converted into a Lightning Memory-Mapped Database<sup>4</sup> (LMDB) before the experiments. The main benefit of LMDB is an extremely fast speed of read transactions, which is important for the training process.

The **IAM** handwriting database [14] contains 1,539 handwritten forms, written by 657 authors, and the total number of words is 115,320, with a lexicon of approximately 13,500 words. This work uses the most widespread RWTH Aachen partition into writer-exclusive training, validation, and test sets [12] to ensure a fair comparison with the related works.

The **Imgur5K** dataset [13] is a new multi-writer dataset that contains 8,094 images, with the handwriting of approximately 5,305 authors published on an online image-sharing and image-hosting service imgur.com. The total number of words is 135,375 with a lexicon of approximately 27,000 words. In comparison with the IAM dataset, the Imgur5K dataset contains 8 times more authors, 1.2 times more words, and 2 times larger lexicon. It is also a more challenging dataset in terms of background clutter and variability in pen types and styles. This work uses the same partitioning into document-exclusive training, validation, and test sets, as was provided from the dataset source. Fig. 2 shows examples of a word *through* in both training and test sets.

The actual size of partitions differs from full partitions if an annotation is missing, the length of the annotation exceeds 25 characters, or if it contains characters that are not included in the model’s character set. Furthermore, the IAM partitions contain only those words where the segmentation of corresponding lines was marked as “OK”. The size of unfiltered and filtered partitions for both case-sensitive and case-insensitive models is shown in Table 1.

### 4.2 Hyper-parameters

AdaDelta optimizer is used with the hyper-parameters matching those in [1]: learning rate  $\eta=1$ ; decay constant  $\rho=0.95$ ; numerical stability term  $\epsilon=1e-8$ ; gradient clipping is 5; maximum number of iterations 300,000. Early stopping patience is set to 5 epochs, and training batch size is 32.

<sup>4</sup> <https://lmdb.readthedocs.io/>

**Table 1.** Size of training, validation, and test partitions used in experiments.

Dataset	Character set	Training	Validation	Testing
IAM	All characters	47,981	7,554	20,305
	Case-sensitive (incl. special characters)	47,963	7,552	20,300
	Case-insensitive	41,228	6,225	17,381
Imgur5K	All characters	182,528	22,470	23,663
	Case-sensitive (incl. special characters)	164,857	19,932	21,509
	Case-insensitive	138,577	16,823	18,110
Total	All characters	230,509	30,024	43,968
	Case-sensitive (incl. special characters)	212,820	27,484	41,809
	Case-insensitive	179,805	23,048	35,491

**Table 2.** Validation set errors for the proposed method.

Char. set	Fine-tuned on	Val-CER			Val-WER		
		Imgur5K	IAM	Both	Imgur5K	IAM	Both
Case-insensitive	Benchmark	22.66	25.65	23.47	37.79	47.05	40.29
	IAM	27.58	3.65	21.12	40.83	11.79	32.98
	IAM→Imgur5K	<b>6.28</b>	7.73	6.67	<b>12.95</b>	20.06	14.87
	Imgur5K	6.70	8.48	7.18	13.48	22.43	15.90
	Imgur5K→IAM	13.69	<b>2.73</b>	10.73	23.43	<b>8.76</b>	19.47
	Imgur5K+IAM	7.17	3.24	<b>6.11</b>	13.79	9.82	<b>12.72</b>
Case-sensitive	Benchmark	29.77	42.81	33.35	49.73	59.07	52.30
	IAM	34.80	5.16	26.65	53.59	12.45	42.28
	IAM→Imgur5K	<b>9.32</b>	25.07	13.65	<b>19.61</b>	34.88	23.81
	Imgur5K	9.18	23.31	13.06	19.74	36.32	24.29
	Imgur5K→IAM	22.43	<b>4.88</b>	17.61	38.48	<b>11.80</b>	31.15
	Imgur5K+IAM	9.48	4.97	<b>8.24</b>	19.69	12.38	<b>17.68</b>

### 4.3 Results

Two standard metrics: character error rate (CER) and word error rate (WER) are used to evaluate the experimental results. CER is defined as a normalized Levenshtein distance:

$$CER = \frac{1}{N} \sum_{i=1}^N D(s_i, \hat{s}_i) / l_i,$$

where  $N$  is the number of annotated images in the set,  $D(\cdot)$  is the Levenshtein distance,  $s_i$  and  $\hat{s}_i$  denote the ground truth text string and the corresponding predicted string, and  $l_i$  is the length of the ground truth string. Given the predictions are evaluated at word level, the WER corresponds to the misclassification

**Table 3.** Validation WER on varying the learning rate  $\eta$  and decay constant  $\rho$  for the case-insensitive benchmark model, fine-tuned on the Imgur5K→IAM sequence. The lowest error is shown in **bold**. Cells with the lowest row-wise error have a grey fill.

Learning rate $\eta$	Decay constant $\rho$		
	0.9	0.95	0.99
1.5	9.69	11.07	11.47
1	9.59	<b>8.76</b>	9.98
0.1	9.27	8.96	8.96
0.01	9.30	9.49	9.03
0.001	11.33	10.86	10.67

error:

$$WER = \frac{1}{N} \sum_{i=1}^N \mathbb{1}\{s_i \neq \hat{s}_i\}.$$

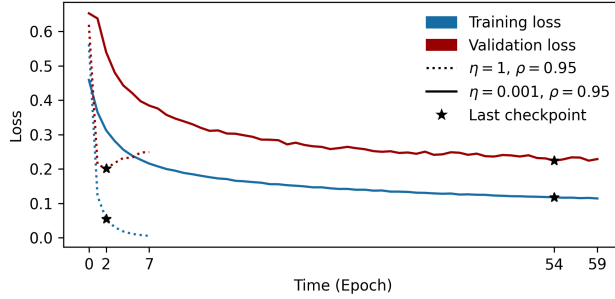
Table 2 presents the validation set CER and WER. Models with a case-sensitive character set have 36 alphanumeric characters and a special EOS token. Models with a case-insensitive character set additionally have 26 upper-case letters and 32 special characters. In all experiments, training and validation sets are i.i.d. (independent and identically distributed). Sequential transfer learning using one and then the other dataset is denoted with a right arrow between the dataset names. Usage of two datasets as a union is denoted with a plus sign.

When transfer learning is performed sequentially, the lowest error on either Imgur5K or IAM is obtained when the corresponding dataset is used last in the fine-tuning sequence. The lowest error on the union of two datasets is obtained when both datasets are used for transfer learning at the same time. These models also have a competitive performance on stand-alone datasets. Furthermore, the results show that the model’s domain of application determines the fine-tuning approach. If the unseen data is close to the IAM distribution then a fine-tuning sequence Imgur5K→IAM is expected to give the lowest error. If the unseen data is close to the Imgur5K distribution then two datasets should be used in a reversed sequence. And a more general-purpose handwritten text recognition model, as represented by a union of the two datasets, is more likely to be obtained by fine-tuning on the union of the datasets.

#### 4.4 Hyper-parameter tuning

AdaDelta optimizer’s learning rate  $\eta$  and decay constant  $\rho$  are tuned for the case-insensitive benchmark model fine-tuned on the Imgur5K→IAM sequence. Table 3 summarizes the WER on the IAM validation set. The lowest error corresponds to the default  $\eta = 1$  and  $\rho = 0.95$ . As the learning rate increases from 0.001 to 1.5, lower error values correspond to lower decay constants. This indicates that large steps in a gradient descent work better when less weight is given to the last gradient. This observation is consistent with [19] that suggests a large





**Fig. 3.** Comparison of training and validation loss for the case-insensitive benchmark model fine-tuned on the Imgur5K→IAM sequence using two sets of hyper-parameters.

momentum (i.e., 0.9-0.99) with a constant learning rate can act as a pseudo increasing learning rate, and speed up the training. Whereas oversized momentum values can cause poor training results as the parameters of the pre-trained model get too distorted.

The IAM dataset is relatively small with only 41K training images, compared to 14.4M training images of the corresponding benchmark model, and it is therefore recommended to perform fine-tuning with low learning rates. However, a low learning rate  $\eta = 0.001$  led to relatively high errors for all values of the decay constant. Fig. 3 highlights that training with  $\eta = 0.001$  was stopped when its training loss (solid blue line) was significantly above the training loss achieved with  $\eta = 1$  (dotted blue line). This indicates an under-fitting problem caused by the early stopping mechanism. By switching the early stopping mechanism off and training for the full 300,000 iterations, we can achieve an error reduction from 10.86%→9.4% in validation WER. In future work, other regularization strategies will be investigated such as the norm penalty or dropout.

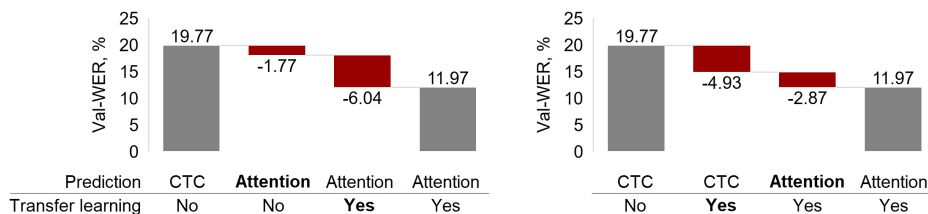
#### 4.5 Ablation study

An ablation study is conducted to demonstrate the effectiveness of transfer learning and attention mechanism for the given problem, and also to analyze the impact on the overall performance on the IAM dataset. This study is performed using a case-sensitive model trained and validated on the IAM dataset, and each model was trained five times using different random seeds. Table 4 presents the average validation set CER and WER. When transfer learning from the STR domain and the Imgur5K dataset is not used, the model’s parameters are initialized randomly (indicated as "None" in the transfer learning column). When the attention mechanism is not used for prediction, it is replaced by the CTC.

The ablation study proves that transfer learning and attention mechanism help reduce both CER and WER. The lowest errors are achieved when both transfer learning and attention mechanism are used. The effect of each component depends on which of them comes first, as illustrated in Fig. 4. Interestingly,

**Table 4.** Ablation study highlighting the importance of transfer learning and attention mechanism. The lowest errors and corresponding components are shown in **bold**.

Transfer learning	Prediction	Val-CER, % Val-WER, %	
None	CTC	7.12	19.77
None	Attention	6.79	18.01
STR, Imgur5K	CTC	5.32	14.84
<b>STR, Imgur5K</b>	<b>Attention</b>	<b>4.84</b>	<b>11.97</b>



**Fig. 4.** The effect of transfer learning and attention mechanism on validation set WER for the IAM dataset when attention mechanism is introduced first (left), and when transfer learning is introduced first (right). Change is highlighted with **bold**.

in both cases, transfer learning is relatively more important among the two studied components in terms of error reduction.

#### 4.6 Test set errors

Table 5 highlights the **test set CER and WER** for our best performing models on each dataset. A comparison is performed with related works on content attention-based seq2seq models, under the same experimental settings using the IAM dataset. The test set WER values suggest that the proposed method outperformed all the state-of-the-art methods [11], [12], [20]. The test set WER for the proposed method is 15.40%. Kang et al. [11] achieved a WER comparable to our work (15.91%) using a content-based attention mechanism. Kang et al. [11] also used location-based attention, achieving a WER of 15.15%, but since it explicitly includes the location information into the attention mechanism, it is not considered in this work. Furthermore, [11] incorporated a language model, unlike the proposed method. Though an RNN decoder can learn the relations between different characters in a word, it is worth investigating the use of language models to improve the performance of the proposed method as future work.

The test set CER results on the other hand suggest that Johannes et al. [15] outperformed all the state-of-the-art methods, with a test set CER of 5.24%. Similar to the proposed work, [15] did not use a language model. However, Kang et al. [11] improved upon their previous work [12] by integrating a language model, and achieved a low error rate (5.79%). The proposed method achieved 6.50% CER and performs comparably with [15] and [11], outperforming the rest of the methods. Based on our error analysis (discussed in detail in the next

**Table 5.** Results comparison with the state-of-the-art content attention-based seq2seq models for handwritten word recognition.

Character set	Test-CER			Test-WER		
	Imgur5K	IAM	Both	Imgur5K	IAM	Both
<b>Ours</b> , case-insensitive	6.46 <sup>a</sup>	4.30 <sup>b</sup>	5.96 <sup>c</sup>	13.58 <sup>a</sup>	12.82 <sup>b</sup>	13.89 <sup>c</sup>
<b>Ours</b> , case-sensitive	9.47 <sup>a</sup>	6.50 <sup>b</sup>	8.59 <sup>c</sup>	20.45 <sup>a</sup>	<b>15.40<sup>b</sup></b>	18.97 <sup>c</sup>
Kang <i>et al.</i> [12]	-	6.88	-	-	17.45	-
Bluche <i>et al.</i> [3]	-	12.60	-	-	-	-
Sueiras <i>et al.</i> [20]	-	8.80	-	-	23.80	-
Chowdhury <i>et al.</i> [5]	-	8.1	-	-	-	-
Johannes <i>et al.</i> [15]	-	<b>5.24</b>	-	-	-	-
Kang <i>et al.</i> [11] <sup>d</sup>	-	5.79	-	-	15.91	-

Fine-tuning approaches: <sup>a</sup> IAM→Imgur5K; <sup>b</sup> Imgur5K→IAM; <sup>c</sup> Imgur5K+IAM.

<sup>d</sup>Also provides the results using location-based attention mechanism, which is not applicable in the considered content-based attention experimental settings.

section), the CER can be further reduced by using data augmentation, language modeling, and a different regularization method.

We also performed tests on the case-insensitive model that achieved 4.3% CER and 12.82% WER on IAM. This work is also the first attempt at using the Imgur5K dataset for handwritten word recognition (to the best of the authors’ knowledge), and therefore no related work exists to compare the performance on the Imgur5K dataset. However, it can be observed from Table 5 that the proposed model has a relatively low test set CER and WER for Imgur5K, and can model a variety of handwritten words from this novel multi-writer dataset.

In summary, the proposed method has the lowest error rate when evaluated at the word level, and the third-lowest error rate when evaluate at the character level. The advantages of using the proposed approach in comparison with [15] and [11] include addressing the problem of training data scarcity, performing transfer learning from STR to HTR, and using Imgur5K dataset with word images from 5000 different writers. This allows the proposed HTR system to also handle examples that are rare due to insufficient annotated data. Our models are publicly available for further fine-tuning or predictions on unseen data.

## 5 Error analysis

This section performs error analysis for our case-insensitive model performing best on the union of Imgur5K and IAM datasets.

### 5.1 Character-level error analysis

To evaluate the model’s predictive performance at the character level, precision and recall are computed from a confusion matrix. It is based on 33,844

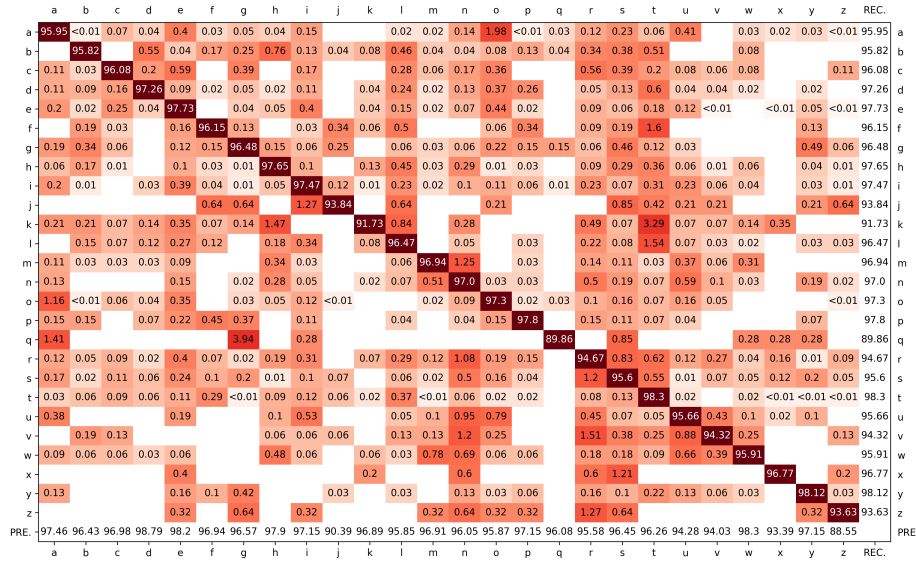
images from 35,491 images in the test set where the length of the predicted word matches the length of the ground truth word. This is necessary to have 1-to-1 correspondence between true and predicted characters. Fig. 5 presents a normalized confusion matrix with the ground truth characters in rows and predicted characters in columns. Only letters are shown for better readability. Values in each row are divided by the total number of characters in that row. Therefore, the main diagonal contains recall values, and each row adds up to 1, possibly with a rounding error. Off-diagonal values represent errors. The recall is also shown on the right margin and precision is shown on the bottom margin. All values are in percent, rounded to two decimal digits. Additionally, precision and recall are combined into a single metric, F1 score, and plotted against the probability mass (normalized frequency) of each character on Fig. 6. These two figures present several interesting and valuable insights:

- The model’s errors pass a sanity check as the confusion matrix highlights that the most typical errors are between characters having similar ways of writing. For example, *a* is most often confused with *o*, and *q* with *g*.
- Letters *z*, *q*, *j* and *x* and all digits are significantly under-represented in the distribution of characters.
- In general, a lower probability mass of a character leads to a lower F1 score. However, the uniqueness of a character’s way of writing seems to be an important factor. For example, digits *0* and *3* have approximately the same probability mass but hugely different F1 scores (77% and 97%). This difference is attributed to the fact that digit *3* is relatively unique and is, therefore, more difficult to confuse with other characters. The same conclusion applies to letter *x*, which achieves a relatively high F1 score despite being among the least frequent characters. Intuitively, characters that are somewhat unique in the way they are written need fewer examples in the training set to achieve high F1 scores. This also suggests that F1 scores can be used to prioritize characters for inclusion in the data augmentation, as a possible next step to improve the model’s performance.

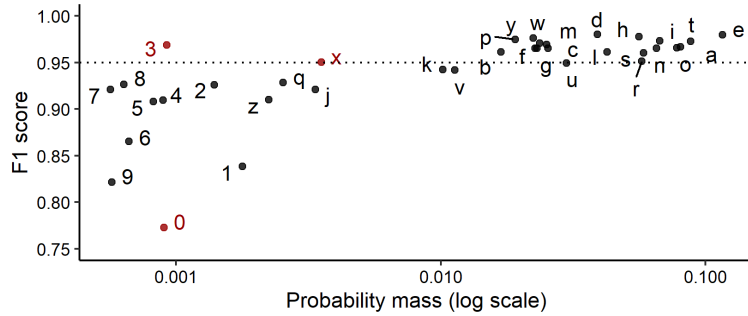
## 5.2 Bias-variance analysis

To gauge the model’s performance, bias-variance analysis is performed according to [16]. It yields only approximate values of the test error components but is helpful and instructive from a practical point of view. Human-level error is used as a proxy for Bayes error, which is analogous to irreducible error and denotes the lowest error that can be achieved on the training set before the model’s parameters start over-fitting. To estimate a human-level error, 22 human subjects from different backgrounds were asked to annotate the same set of randomly selected 100 misclassified images from the test set. The images that are misclassified by *all* reviewers contribute to the human-level error estimate.

In general, the bias of the model is the difference between Bayes error and the training set error. The variance is the difference between the training and

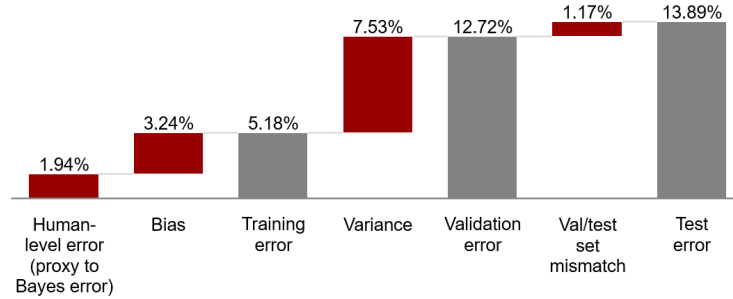


**Fig. 5.** Character-level error analysis using a confusion matrix with a row-wise normalization. The main diagonal represents recall. Each row adds up to 1, possibly with a rounding error. Abbreviations used: PRE. stands for precision, REC. stands for recall.



**Fig. 6.** F1 scores vs probability mass of lower-case alphanumeric characters. Characters discussed in the text are depicted in red. Figure best viewed in color.

validation set error. Additionally, the difference between the validation and test set error is denoted by the validation/test set mismatch error. Fig. 7 summarizes the results of the bias-variance analysis. By definition, it is not feasible to reduce the Bayes error, which is 1.94% according to our estimate. It can be observed in Fig. 7 that the variance contributes the largest part of the error (7.53%) among other components and should be the primary focus of measures aimed at decreasing the testing error of the model. Suitable strategies include using a larger training set, data augmentation, and a different regularization method



**Fig. 7.** Bias-variance decomposition. Red bars are components of the test set WER on the union of Imgur5K and IAM. Grey bars are the model’s errors obtained by summing red bars from left to right. Human-level error is used as a proxy to Bayes error.

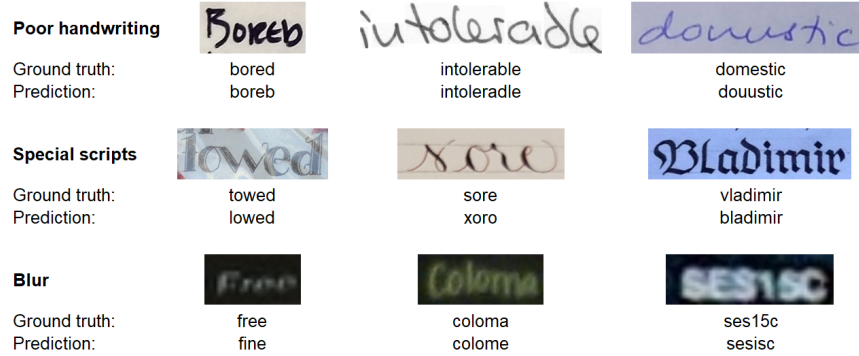
(e.g. norm penalty, dropout) instead of early stopping. The second largest component is the bias (3.24%) that can be reduced by, for example, training a bigger neural network, using a different optimizer (e.g. Adam), and running the training longer. The smallest component (1.17%) is due to the validation/test set mismatch; this can be addressed by reviewing the partitioning strategy for the IAM dataset. However, this work follows the same splits as widely used in the HTR research for a fair comparison with other related work.

### 5.3 Visual analysis of images

To provide further insights into the possible causes for the model’s errors, a visual inspection is performed on the same set of 100 misclassified images. The main problems contributing towards errors include poor handwriting, the use of special scripts (e.g. Gothic, curlicue, etc.), errors in some ground truth images, and other forms of issues such as low image resolution, blurry and rotated text, challenging backgrounds, strike-through words, or multiple words in an image. Fig. 8 contains sample images representing some of the listed problems. The analysis suggests that the performance of the model can be further improved using data augmentation (with special scripts, blur, rotations, variety of backgrounds, etc.), and also fine-tuning on other multi-writer datasets. A language model is also foreseen to reduce the error rate on images with poor handwriting.

## 6 Conclusion

This work presented an end-to-end-HTR system based on attention encoder-decoder networks, which leverages pre-trained models trained on a large set of synthetic scene text images for handwritten word recognition. The problem of training data scarcity is addressed by performing transfer learning from STR to HTR, and using a new multi-writer dataset (Imgur5K) that contains word examples from 5000 different writers. This allows the proposed HTR system



**Fig. 8.** Sample images from selected categories. Each category represents a possible main cause for misclassification, as identified through visual analysis.

to also cope with rare examples due to insufficient annotated data. The experimental results on multi-writer datasets (IAM and Imgur5K) demonstrate the effectiveness of the proposed method. Under the given experimental settings, the proposed method outperformed the state-of-the-art methods by achieving the lowest error rate when evaluated at the word level on the IAM dataset (test set WER 15.40%). At character level, the proposed method performed comparable with the state-of-the-art methods and achieved 6.50% test set CER. However, the character level error can be further reduced by using data augmentation, language modeling, and a different regularization method, which will be investigated as future work. Our source code and pre-trained models are publicly available for further fine-tuning or predictions on unseen data at GitHub<sup>5</sup>.

## Acknowledgment

The authors would like to thank the Centre for Digital Humanities Uppsala, SSBA, Anders Hast and Örjan Simonsson for their kind support and encouragement provided during the project. The computations were performed on resources provided by SNIC through Alvis @ C3SE under project SNIC 2021/7-47.

## References

1. Baek, J., Kim, G., Lee, J., Park, S., Han, D., Yun, S., Oh, S.J., Lee, H.: What is wrong with scene text recognition model comparisons? dataset and model analysis. In: IEEE International Conference on Computer Vision. pp. 4715–4723 (2019)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
3. Bluche, T., Louradour, J., Messina, R.: Scan, attend and read: End-to-end handwritten paragraph recognition with mdlstm attention. In: 14th IAPR international conference on document analysis and recognition. vol. 1, pp. 1050–1055 (2017)

<sup>5</sup> <https://github.com/dmitrijsk/AttentionHTR>

4. Bookstein, F.L.: Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on pattern analysis and machine intelligence* **11**(6), 567–585 (1989)
5. Chowdhury, A., Vig, L.: An efficient end-to-end neural model for handwritten text recognition. *arXiv preprint arXiv:1807.07965* (2018)
6. Dutta, K., Krishnan, P., Mathew, M., Jawahar, C.: Improving cnn-rnn hybrid networks for handwriting recognition. In: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 80–85. IEEE (2018)
7. Frinken, V., Fischer, A., Manmatha, R., Bunke, H.: A novel word spotting method based on recurrent neural networks. *IEEE transactions on pattern analysis and machine intelligence* **34**(2), 211–224 (2012)
8. Gupta, A., Vedaldi, A., Zisserman, A.: Synthetic data for text localisation in natural images. In: *IEEE conference on computer vision and pattern recognition*. pp. 2315–2324 (2016)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *IEEE conference on computer vision and pattern recognition*. pp. 770–778 (2016)
10. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227* (2014)
11. Kang, L., Riba, P., Villegas, M., Fornés, A., Rusiñol, M.: Candidate fusion: Integrating language modelling into a sequence-to-sequence handwritten word recognition architecture. *Pattern Recognition* **112**, 107790 (2021)
12. Kang, L., Toledo, J.I., Riba, P., Villegas, M., Fornés, A., Rusiñol, M.: Convoive, attend and spell: An attention-based sequence-to-sequence model for handwritten word recognition. In: *German Conference on Pattern Recognition*. pp. 459–472. Springer (2018)
13. Krishnan, P., Kovvuri, R., Pang, G., Vassilev, B., Hassner, T.: Textstylebrush: Transfer of text aesthetics from a single example. *arXiv preprint arXiv:2106.08385* (2021)
14. 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)
15. Michael, J., Labahn, R., Grüning, T., Zöllner, J.: Evaluating sequence-to-sequence models for handwritten text recognition. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1286–1293. IEEE (2019)
16. Ng, A.: Lecture notes in cs230 deep learning (Stanford University, 2021 Fall)
17. Rodríguez-Serrano, J.A., Perronnin, F., et al.: A model-based sequence similarity with application to handwritten word spotting. *IEEE transactions on pattern analysis and machine intelligence* **34**(11), 2108–2120 (2012)
18. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence* **39**(11), 2298–2304 (2016)
19. Smith, L.N.: A disciplined approach to neural network hyper-parameters: Part 1—learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820* (2018)
20. Sueiras, J., Ruiz, V., Sanchez, A., Velez, J.F.: Offline continuous handwriting recognition using sequence to sequence neural networks. *Neurocomputing* **289**, 119–128 (2018)
21. Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., Liu, C.: A survey on deep transfer learning. In: *International conference on artificial neural networks*. pp. 270–279. Springer (2018)