An attention based method for offline handwritten Urdu text recognition

Tayaba Anjum and **Nazar Khan**

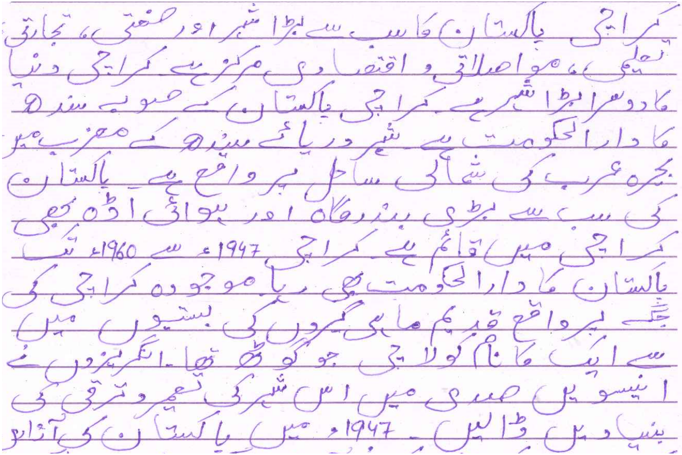
Computer Vision and Machine Learning Group,
Punjab University College of Information Technology (PUCIT)
University of the Punjab, Lahore, Pakistan

17th International Conference on Frontiers in Handwriting Recognition
September 9, 2020

Problem: Offline Handwritten Text Recognition (OHTR) for Urdu

Offline Handwritten Text Recognition (OHTR)

Automatic recognition of text from images of handwritten text.



Urdu is a right-to-left language written in a mixture of mostly cursive and occasionally non-cursive form.

Why is handwritten Urdu recognition difficult?

- Sayre's paradox: cursively written word cannot be recognized without being segmented and cannot be segmented without being recognized.
- Wide range of variations in pen type, writing style, writing size, page background.
- Almost every rule can be violated until text becomes illegible.

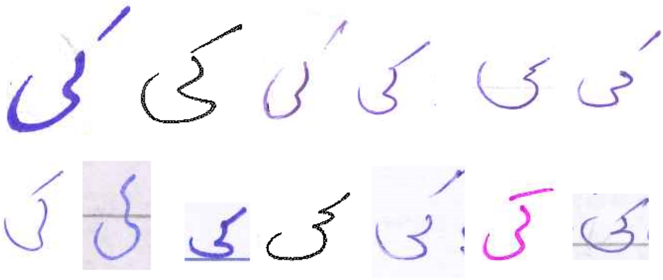


Figure: An Urdu word pronounced as "key" written by 13 different people.

Handwritten vs. Typed Urdu vs. Characters

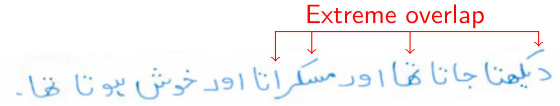
دیکھتا جاتا تھا اور مسکراتا اور خوش ہوتا تھا۔

دیکھتا جاتا تھا اور مسکراتا اور خوش ہوتا تھا

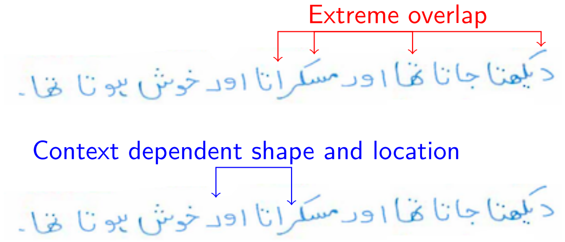
د ی ک ھ ت ا ج ا ت ا ت ھ ا ا و ر م س ک ر ا ت ا ا و ر خ و ش ھ و ت ا ت ھ ا

- Within cursive words, characters to be recognized often appear with widely varying ligatures.
- While typed Urdu conforms to some rules, handwriting can be restricted to very few rules.

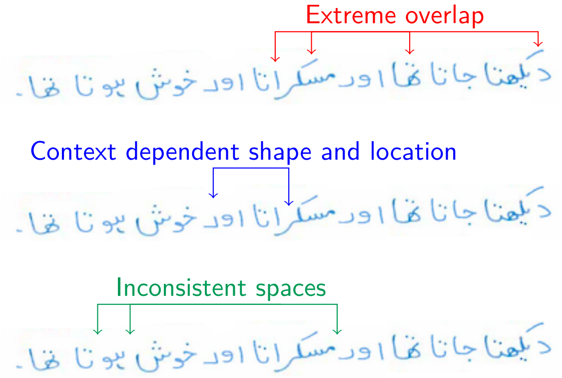
Challenges of Urdu OHTR



Challenges of Urdu OHTR



Challenges of Urdu OHTR



State of OHTR for Arabic-like scripts

Techniques

- Urdu
  - Raw pixels + Bidirectional LSTM[1]
  - CNN + Bidirectional LSTM[2]
- Arabic
  - CNN + LSTM with context-window[3]
  - Raw pixels + Multidimensional LSTM[4]
- Similar situation for Farsi.
- The concept of **attention has not been explored**.
- Error reporting has been restricted to the level of characters. **World level accuracy is not reported.**

[1] Saad Bin Ahmed et al. "Handwritten Urdu character recognition using one-dimensional BLSTM classifier". In: *Neural Computing and Applications* 31.4 (2019), pp. 1143–1151.

[2] S. Hassan et al. "Cursive Handwritten Text Recognition using Bi-Directional LSTMs: A Case Study on Urdu Handwriting". In: *International Conference on Deep Learning and Machine Learning in Emerging Applications, Deep-ML 2019*. 2019.

[3] Gui et al. "Adaptive context-aware reinforced agent for handwritten text recognition". In: *Proceedings of the British Machine Vision Conference (BMVC)*. 2018.

[4] Riaz Ahmad et al. "KHATT: A deep learning benchmark on Arabic script". In: *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2017. DOI: 10.1109/ICDAR.2017.358.

State of OHTR for Arabic-like scripts
Datasets

- Urdu
  - CENIP-UCCP[5]
  - UCOM/UNHD[6]
  - Hassan et al.[7]
- Arabic
  - IFN/ENIT[8]
  - OpenHaRT[9]
  - KHATT[10]
- Farsi
  - Sadri et al.[11]
- Urdu datasets are **not entirely accessible**.

[5] Raza et al., "An Unconstrained Benchmark Urdu Handwritten Sentence Database with Automatic Line Segmentation".

[6] Ahmed et al., "Handwritten Urdu character recognition using one-dimensional BLSTM classifier".

[7] Hassan et al., "Cursive Handwritten Text Recognition using Bi-Directional LSTMs: A Case Study on Urdu Handwriting".

[8] Pechwitz et al., "IFN/ENIT-database of handwritten Arabic words".

[9] Tong et al., "NIST 2013 open handwriting recognition and translation evaluation".

[10] Mahmoud et al., "KHATT: An open Arabic offline handwritten text database".

[11] Sadri, Yeganehzad, and Saghi, "A novel comprehensive database for offline Persian handwriting recognition".

Contributions

Paper Contributions

1. We present the first *attention-based*, encoder-decoder model for recognition of offline handwritten Urdu text.

Paper Contributions

1. We present the first *attention-based*, encoder-decoder model for recognition of offline handwritten Urdu text.
2. We present a new dataset of offline handwritten Urdu text containing $7,309$ unique text line images with ground-truth.

Paper Contributions

1. We present the first *attention-based*, encoder-decoder model for recognition of offline handwritten Urdu text.
2. We present a new dataset of offline handwritten Urdu text containing $7,309$ unique text line images with ground-truth.
3. For the first time, we report word level accuracy instead of only character level accuracy.

Paper Contributions

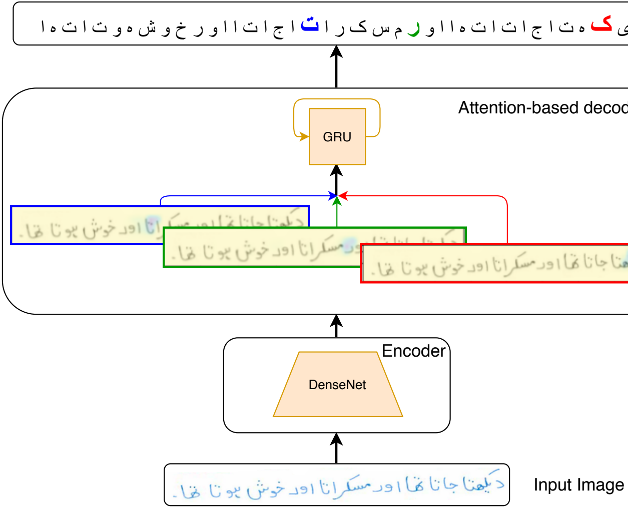
1. We present the first *attention-based*, encoder-decoder model for recognition of offline handwritten Urdu text.
2. We present a new dataset of offline handwritten Urdu text containing $7,309$ unique text line images with ground-truth.
3. For the first time, we report word level accuracy instead of only character level accuracy.
4. We show that attention can enable uni-directional decoders to out-perform bidirectional decoders.

Paper Contributions

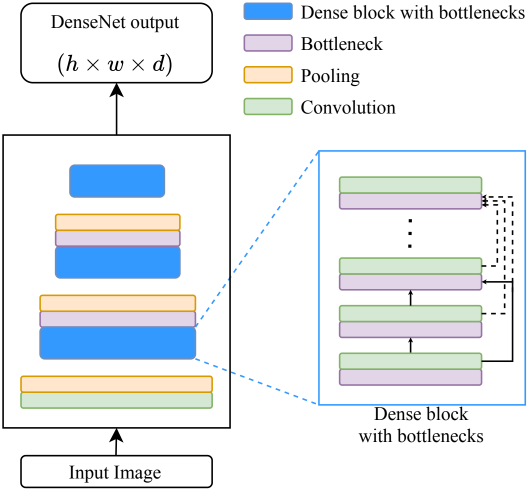
1. We present the first *attention-based*, encoder-decoder model for recognition of offline handwritten Urdu text.
2. We present a new dataset of offline handwritten Urdu text containing $7,309$ unique text line images with ground-truth.
3. For the first time, we report word level accuracy instead of only character level accuracy.
4. We show that attention can enable uni-directional decoders to out-perform bidirectional decoders.
5. Compared to previous approaches, we report close to 2x accuracy improvement at character level and 37x at word level.

Proposed Solution

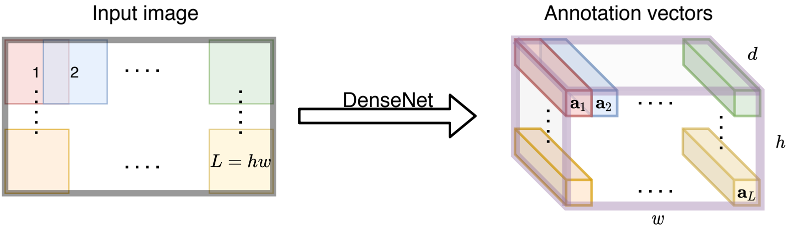
Encoder/Decoder Model



Encoder: DenseNet with bottlenecks



Interpretation of DenseNet Output



Output volume of size $h \times w \times d$ represents d-dimensional *annotation vectors* of $h \times w$ overlapping blocks of the input image.

Decoder: Gated Recurrent Unit (GRU)

Produces text sequence one character at a time conditioned on the previously generated character $\mathbf{y}_{t-1}$ and current hidden state $\mathbf{h}_t$.

$$\begin{aligned}
\mathbf{z}_t &= \sigma(\mathbf{W}_{yz}\mathbf{E}\mathbf{y}_{t-1} + \mathbf{U}_{hz}\mathbf{h}_{t-1}) \\
\mathbf{r}_t &= \sigma(\mathbf{W}_{yr}\mathbf{E}\mathbf{y}_{t-1} + \mathbf{U}_{hr}\mathbf{h}_{t-1}) \\
\widetilde{\mathbf{h}}_t &= \tanh(\mathbf{W}_{yh}\mathbf{E}\mathbf{y}_{t-1} + \mathbf{U}_{rh}(\mathbf{r}_t \otimes \mathbf{h}_{t-1})) \\
\mathbf{h}_t &= (1 - \mathbf{z}_t) \otimes \mathbf{h}_{t-1} + \mathbf{z}_t \otimes \widetilde{\mathbf{h}}_t
\end{aligned}$$

where update gate $\mathbf{z}_t$ and rest gate $\mathbf{r}_t$ and candidate hidden state $\widetilde{\mathbf{h}}_t$ are used to compute the hidden state and GRU output $\mathbf{h}_t$.

Decoder: Gated Recurrent Unit (GRU) with Attention

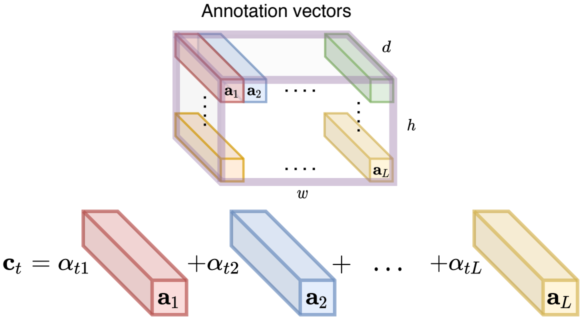
Produces text sequence one character at a time conditioned on the previously generated character $\mathbf{y}_{t-1}$, current hidden state $\mathbf{h}_t$ and a context vector $\mathbf{c}_t$.

$$
\begin{aligned}
\mathbf{z}_t &= \sigma(\mathbf{W}_{yz}\mathbf{E}\mathbf{y}_{t-1} + \mathbf{U}_{hz}\mathbf{h}_{t-1} + \mathbf{C}_{cz}\mathbf{c}_t) \\
\mathbf{r}_t &= \sigma(\mathbf{W}_{yr}\mathbf{E}\mathbf{y}_{t-1} + \mathbf{U}_{hr}\mathbf{h}_{t-1} + \mathbf{C}_{cr}\mathbf{c}_t) \\
\widetilde{\mathbf{h}}_t &= \tanh(\mathbf{W}_{yh}\mathbf{E}\mathbf{y}_{t-1} + \mathbf{U}_{rh}(\mathbf{r}_t \otimes \mathbf{h}_{t-1}) + \mathbf{C}_{ch}\mathbf{c}_t) \\
\mathbf{h}_t &= (1 - \mathbf{z}_t) \otimes \mathbf{h}_{t-1} + \mathbf{z}_t \otimes \widetilde{\mathbf{h}}_t
\end{aligned}
$$

where update gate $\mathbf{z}_t$ and rest gate $\mathbf{r}_t$ and candidate hidden state $\widetilde{\mathbf{h}}_t$ are used to compute the hidden state and GRU output $\mathbf{h}_t$.

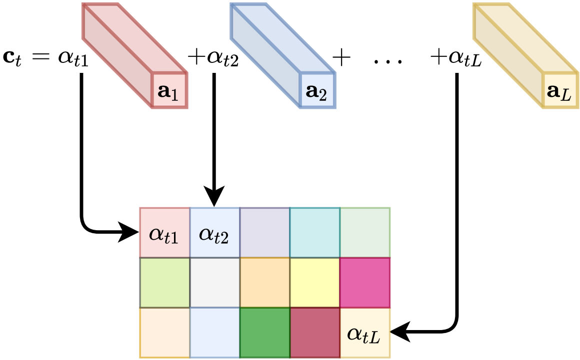
Attention via the context vector

The context vector $\mathbf{c}_t$ is a dynamic representation of the relevant part of the image at time t. Computed as a weighted sum of annotation vectors



Attention via the context vector

Attention weight α_{ti} determines importance of location i in determining context vector $\mathbf{c}_t$.



Attention weights

Standard model

Attention weights α_{ti} computed via softmax over locations

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{j=1}^{L} \exp(e_{tj})}$$

Per-location exponents computed via

$$e_{ti} = \mathbf{v}_a^T \tanh(\mathbf{W}_a \mathbf{h}_{t-1} + \mathbf{U}_a \mathbf{a}_i)$$

depend on

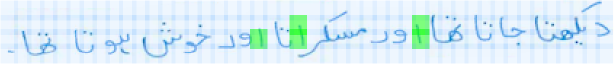
- the hidden state, and
- each location's content encoded in the annotation vector

Computing Attention

- Ideally, we would like our text recognizer to “read” like we do.
- It should focus on the relevant image region when recognizing each character.

Computing Attention

- Ideally, we would like our text recognizer to “read” like we do.
- It should focus on the relevant image region when recognizing each character.



- However, a character or word can appear at multiple locations.
- Nothing stops an attention model from re-attending a previously attended location.
- The decision for attention *in text* needs to depend on the history of attention.

Computing Attention with Coverage

Proposed model

Attention weights α_{ti} computed via softmax over locations

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{j=1}^{L} \exp(e_{tj})}$$

Exponents computed via

$$e_{ti} = \mathbf{v}_a^T \tanh(\mathbf{W}_a \mathbf{h}_{t-1} + \mathbf{U}_a \mathbf{a}_i + \mathbf{U}_f \mathbf{f}_i)$$

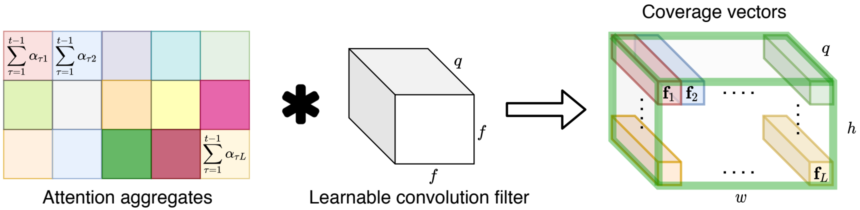
where *coverage vector*[12] $\mathbf{f}_i$ represents history of attention already given to location i.

[12] Zhang, Du, and Dai, "Multi-scale attention with dense encoder for handwritten mathematical expression recognition".

Coverage vectors

2-step computation of coverage vectors $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_L$ at time t

1. Compute aggregated attention $\sum_{\tau=1}^{t-1} \alpha_{\tau i}$ at each location i.
2. Convolve attention aggregates with q filters of size $f \times f$.



Attention and coverage

Attention of location i

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{j=1}^{L} \exp(e_{tj})}$$

$$e_{ti} = \mathbf{v}_a^T \tanh(\mathbf{W}_a \mathbf{h}_{t-1} + \mathbf{U}_a \mathbf{a}_i + \mathbf{U}_f \mathbf{f}_i)$$

depends on

- hidden state,
- description of location i, and
- attention already given to location i.

Output Character Probability

Conditional probabilities of next character computed via softmax

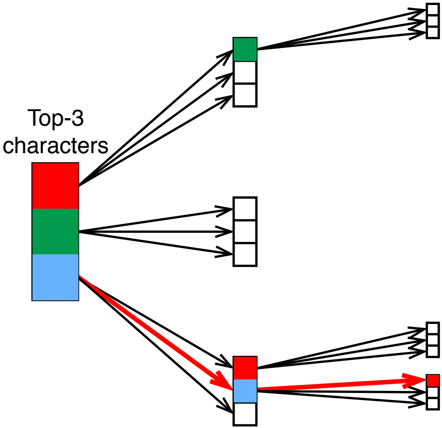
$$p(\mathbf{y}_t|\mathbf{y}_{t-1};\mathbf{X}) = \text{softmax}(\mathbf{W}_o(\mathbf{E}\mathbf{y}_{t-1} + \mathbf{W}_h\mathbf{h}_t + \mathbf{W}_c\mathbf{c}_t))$$

depend on

- previous output character $\mathbf{y}_{t-1}$,
- hidden state $\mathbf{h}_t$, and
- context $\mathbf{c}_t$ which
  - encodes localized image region(s),
  - based on a history of localizations/coverage.

Optimal character sequence

Optimal sequence of characters $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_c$ is found via beam search.



$t = 1$ $t = 2$ $t = 3$

New Dataset

New Dataset

Statistics

- 7,309 offline handwritten Urdu text lines.
- 78,870 words written by 100 different writers.
- 98 unique Urdu characters.

Collection

- 100 undergraduate students between the ages 20-24 years.
- Requested to submit
  - any handwritten Urdu text, and
  - a corresponding ground-truth text file.
- No restriction on pen type, page type, and ink colour.
- No restriction on what to write.
- Pages scanned at 200 DPI and text lines manually segmented.
- No deskewing.
- Submitted ground-truth was thoroughly checked and corrected/completed by a team of 3 persons.

Comparison of datasets

Dataset	Total Lines	Total Words	Total characters	Total Writers	Vocabulary Size
UCOM/UNHD[13]	10,000	312,000	1,872,000	500	59
CENIP-UCCP[14]	2,051	23,833	-	200	-
Hassan et al.[15]	6,000	86,400	432,000	600	-
Proposed	**7,309**	**78,870**	**283,664**	**100**	**98**

[13] Saad Bin Ahmed et al. "Handwritten Urdu character recognition using one-dimensional BLSTM classifier". In: *Neural Computing and Applications* 31.4 (2019), pp. 1143–1151.

[14] A. Raza et al. "An Unconstrained Benchmark Urdu Handwritten Sentence Database with Automatic Line Segmentation". In: *International Conference on Frontiers in Handwriting Recognition*. 2012, pp. 491–496.

[15] S. Hassan et al. "Cursive Handwritten Text Recognition using Bi-Directional LSTMs: A Case Study on Urdu Handwriting". In: *International Conference on Deep Learning and Machine Learning in Emerging Applications, Deep-ML 2019*. 2019.

Comparison of datasets

Dataset	Total Lines	Total Words	Total characters	Total Writers	Vocabulary Size
UCOM/UNHD[13]	10,000	312,000	1,872,000	500	59
CENIP-UCCP[14]	2,051	23,833	-	200	-
Hassan et al.[15]	6,000	86,400	432,000	600	-
Proposed	**7,309**	**78,870**	**283,664**	**100**	**98**

Existing datasets:

- Unreliable statistics. Only 700 lines in UCOM/UNHD are unique.
- Either publicly unavailable or only partially available.

[13] Saad Bin Ahmed et al. "Handwritten Urdu character recognition using one-dimensional BLSTM classifier". In: *Neural Computing and Applications* 31.4 (2019), pp. 1143–1151.

[14] A. Raza et al. "An Unconstrained Benchmark Urdu Handwritten Sentence Database with Automatic Line Segmentation". In: *International Conference on Frontiers in Handwriting Recognition*. 2012, pp. 491–496.

[15] S. Hassan et al. "Cursive Handwritten Text Recognition using Bi-Directional LSTMs: A Case Study on Urdu Handwriting". In: *International Conference on Deep Learning and Machine Learning in Emerging Applications, Deep-ML 2019*. 2019.

Comparison of datasets

Dataset	Total Lines	Total Words	Total characters	Total Writers	Vocabulary Size
UCOM/UNHD[13]	10,000	312,000	1,872,000	500	59
CENIP-UCCP[14]	2,051	23,833	-	200	-
Hassan et al.[15]	6,000	86,400	432,000	600	-
Proposed	**7,309**	**78,870**	**283,664**	**100**	**98**

Existing datasets:

- Unreliable statistics. Only 700 lines in UCOM/UNHD are unique.
- Either publicly unavailable or only partially available.

Proposed dataset publicly available in full at

`http://faculty.pucit.edu.pk/nazarkhan/`

[13] Saad Bin Ahmed et al. "Handwritten Urdu character recognition using one-dimensional BLSTM classifier". In: *Neural Computing and Applications* 31.4 (2019), pp. 1143–1151.

[14] A. Raza et al. "An Unconstrained Benchmark Urdu Handwritten Sentence Database with Automatic Line Segmentation". In: *International Conference on Frontiers in Handwriting Recognition*. 2012, pp. 491–496.

[15] S. Hassan et al. "Cursive Handwritten Text Recognition using Bi-Directional LSTMs: A Case Study on Urdu Handwriting". In: *International Conference on Deep Learning and Machine Learning in Emerging Applications, Deep-ML 2019*. 2019.

Experiments and Results

Attention Visualization

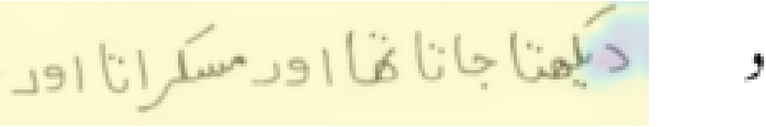


Figure: Our model implicitly segments and recognizes every character by focusing only on localized, relevant areas. It also learns to focus in context.

Attention Visualization

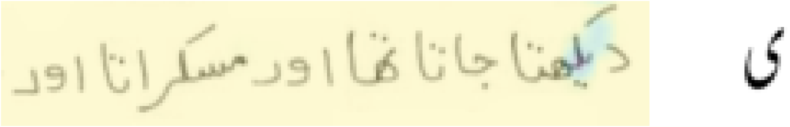


Figure: Our model implicitly segments and recognizes every character by focusing only on localized, relevant areas. It also learns to focus in context.

Attention Visualization

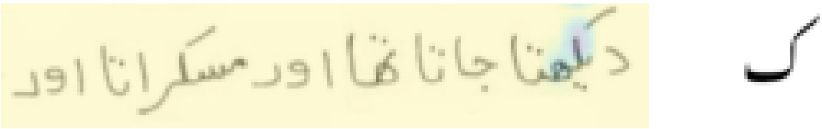


Figure: Our model implicitly segments and recognizes every character by focusing only on localized, relevant areas. It also learns to focus in context.

Attention Visualization

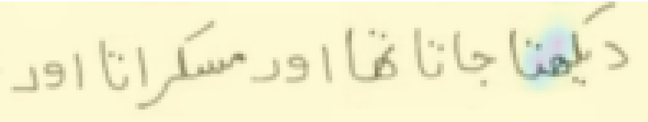


Figure: Our model implicitly segments and recognizes every character by focusing only on localized, relevant areas. It also learns to focus in context.

Attention Visualization

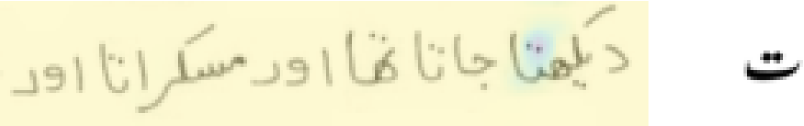


Figure: Our model implicitly segments and recognizes every character by focusing only on localized, relevant areas. It also learns to focus in context.

Attention Visualization

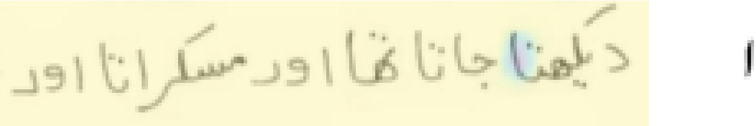


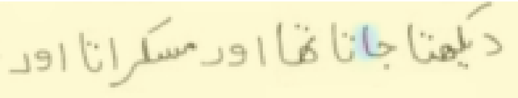
Figure: Our model implicitly segments and recognizes every character by focusing only on localized, relevant areas. It also learns to focus in context.

Attention Visualization

Figure: Our model implicitly segments and recognizes every character by focusing only on localized, relevant areas. It also learns to focus in context.


دیکھتا جاتا تھا اور مسکراتا اور

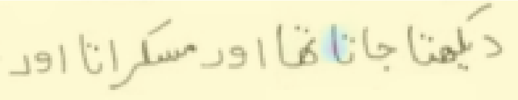
ج


دیکھتا جاتا تھا اور مسکراتا اور

ا


دیکھتا جاتا تھا اور مسکراتا اور

ت


دیکھتا جاتا تھا اور مسکراتا اور

ا


دیکھتا جاتا تھا اور مسکراتا اور

و
ی
ک
ھ
ت
ا
ج
ا
ت
ا
ت
ھ
ا
ا
و
ر
م
س
ک
ر
ا
ت
ا
ا
و
ر

Evaluation Metrics

Character Level

$$\text{CER} = \frac{\text{ins} + \text{sub} + \text{del}}{\text{n}} \times 100$$

Percentage of insertions, substitutions and deletions required to transform target of length n into the output.

$$\text{CLA} = 100 - \text{CER}$$

Word Level

Same formulae as above but with role of characters replaced by words.

Comparison of different models

Models	CLA	WLA
CNN+GRU	39.79	1.24
CNN+LSTM	39.26	1.40
CNN+BGRU	41.45	1.41
CNN+BLSTM[16]	40.90	1.17
DenseNet+GRU+ Attention	**77.05**	**43.35**

[16] S. Hassan et al. "Cursive Handwritten Text Recognition using Bi-Directional LSTMs: A Case Study on Urdu Handwriting". In: *International Conference on Deep Learning and Machine Learning in Emerging Applications, Deep-ML 2019*. 2019.

Comparison of different models

Models	CLA	WLA
CNN+GRU	39.79	1.24
CNN+LSTM	39.26	1.40
CNN+BGRU	41.45	1.41
CNN+BLSTM[16]	40.90	1.17
DenseNet+GRU+ Attention	**77.05**	**43.35**

Reasons for improvement

- More diverse features learned by DenseNet compared to CNN.
- Attention reduces the need for bidirectional decoding.
- Coverage captures right-to-left nature of Urdu.

[16] S. Hassan et al. "Cursive Handwritten Text Recognition using Bi-Directional LSTMs: A Case Study on Urdu Handwriting". In: *International Conference on Deep Learning and Machine Learning in Emerging Applications, Deep-ML 2019*. 2019.

Conclusion

- We have proposed an encoder/decoder framework for recognition of offline handwritten Urdu text.

Conclusion

- We have proposed an encoder/decoder framework for recognition of offline handwritten Urdu text.
- Using attention in context learns the right-to-left nature of Urdu.

Conclusion

- We have proposed an encoder/decoder framework for recognition of offline handwritten Urdu text.
- Using attention in context learns the right-to-left nature of Urdu.
- Close to 2x improvement in character level accuracy.

Conclusion

- We have proposed an encoder/decoder framework for recognition of offline handwritten Urdu text.
- Using attention in context learns the right-to-left nature of Urdu.
- Close to 2x improvement in character level accuracy.
- Around 37x improvement in word level accuracy.

Conclusion

- We have proposed an encoder/decoder framework for recognition of offline handwritten Urdu text.
- Using attention in context learns the right-to-left nature of Urdu.
- Close to 2x improvement in character level accuracy.
- Around 37x improvement in word level accuracy.
- New publicly accessible dataset of >7K text lines consisting of around 79K words written by 100 writers.

`http://faculty.pucit.edu.pk/nazarkhan/`

Conclusion

- We have proposed an encoder/decoder framework for recognition of offline handwritten Urdu text.
- Using attention in context learns the right-to-left nature of Urdu.
- Close to 2x improvement in character level accuracy.
- Around 37x improvement in word level accuracy.
- New publicly accessible dataset of >7K text lines consisting of around 79K words written by 100 writers.

`http://faculty.pucit.edu.pk/nazarkhan/`

Thank you for your attention.

References I

Ahmad, Riaz et al. "KHATT: A deep learning benchmark on Arabic script". In: *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2017. DOI: 10.1109/ICDAR.2017.358.

Ahmed, Saad Bin et al. "Handwritten Urdu character recognition using one-dimensional BLSTM classifier". In: *Neural Computing and Applications* 31.4 (2019), pp. 1143–1151.

Gui et al. "Adaptive context-aware reinforced agent for handwritten text recognition". In: *Proceedings of the British Machine Vision Conference (BMVC)*. 2018.

Hassan, S. et al. "Cursive Handwritten Text Recognition using Bi-Directional LSTMs: A Case Study on Urdu Handwriting". In: *International Conference on Deep Learning and Machine Learning in Emerging Applications, Deep-ML 2019*. 2019.

Mahmoud, Sabri A. et al. "KHATT: An open Arabic offline handwritten text database". In: *Pattern Recognition* 47.3 (2014), pp. 1096–1112.

References II

Pechwitz, Mario et al. "IFN/ENIT-database of handwritten Arabic words". In: *Proc. of CIFED*. Vol. 2. 2002, pp. 127–136.

Raza, A. et al. "An Unconstrained Benchmark Urdu Handwritten Sentence Database with Automatic Line Segmentation". In: *International Conference on Frontiers in Handwriting Recognition*. 2012, pp. 491–496.

Sadri, Javad, Mohammad Reza Yeganehzad, and Javad Saghi. "A novel comprehensive database for offline Persian handwriting recognition". In: *Pattern Recognition* 60 (2016), pp. 378–393.

Tong, A. et al. "NIST 2013 open handwriting recognition and translation evaluation". In: *Proc. NIST*. 2013.

Zhang, Jianshu, Jun Du, and Lirong Dai. "Multi-scale attention with dense encoder for handwritten mathematical expression recognition". In: *2018 24th international conference on pattern recognition (ICPR)*. IEEE. 2018, pp. 2245–2250.