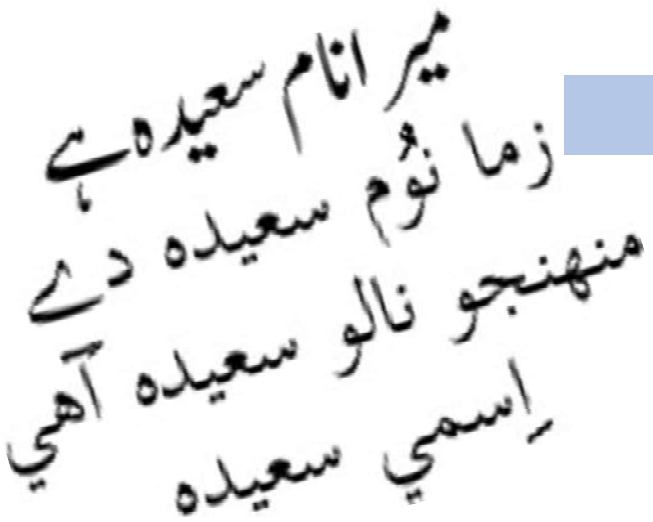
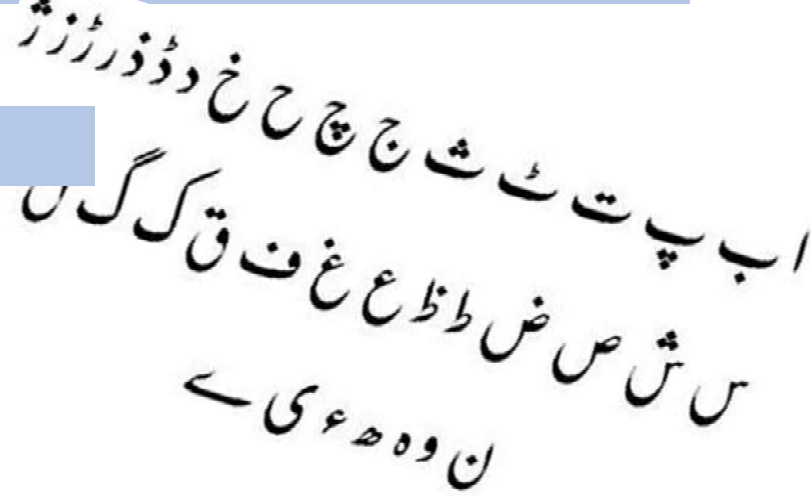
Urdu OCR and Deep Learning: Current and Future





SAEEDA NAZ
سعیدہ ناز

Computer Science Department, Govt. Girls Postgraduate College No.1, Abbottabad

Nastaliq and Nask

Text Recognition

Statistical Machine Learning

Urdu

Pashto

Sindhi

OCR

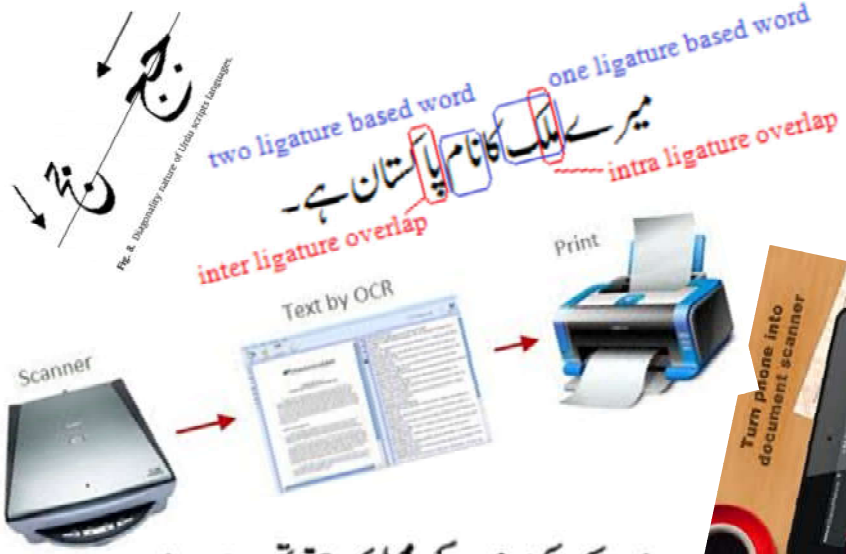
Mobile OCR

Wild Text

Number Plate

Hidden Markov Model

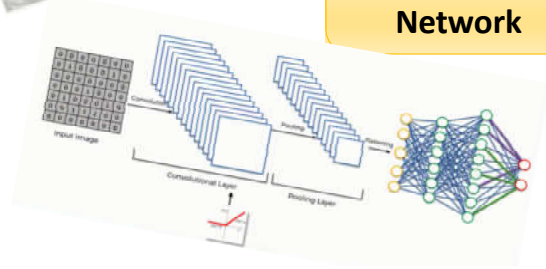
Support Vector Machines

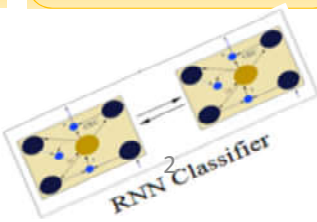


Deep Learning

Scratched Deep Neural Network

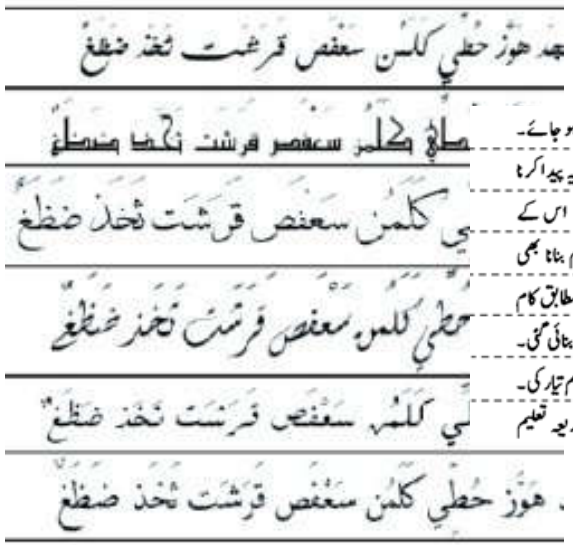
Transfer Learning
-Fine-tuned
-Freeze

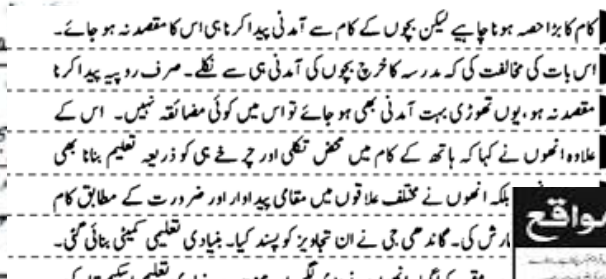




Urdu OCR and its Challenges

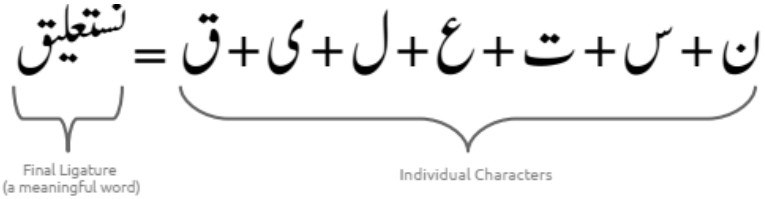
- OCR considered a solved problem???



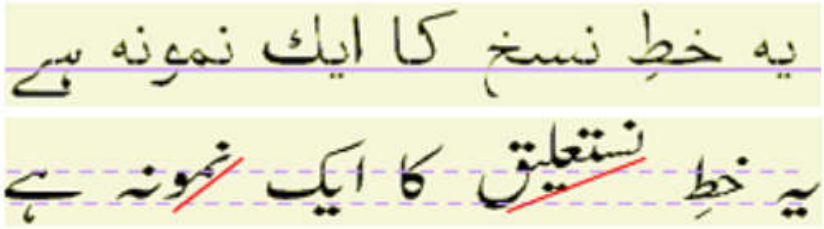


Challenges

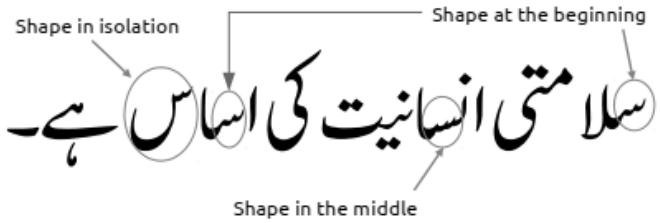
- **Nasta'liq Script**
- **More than 26,000 unique ligature**



A word in Urdu with its constituent characters



Diagonal nature, baseline and height

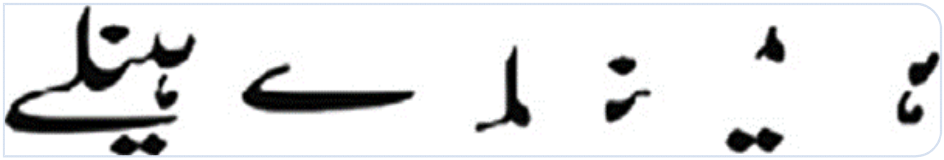


Context shape change

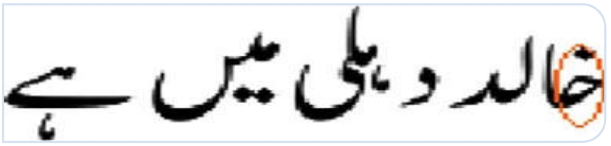
G. S. Lehal, "Choice of recognizable units for urdu ocr," in *Proceeding of the workshop on document analysis and recognition*, 79–85, ACM (2012)

Challenges

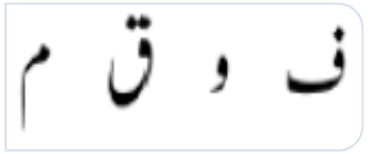
Intra or Inter Ligature(s) overlap



Segmentation Point and thick and thin strokes



False loop



Filled loop

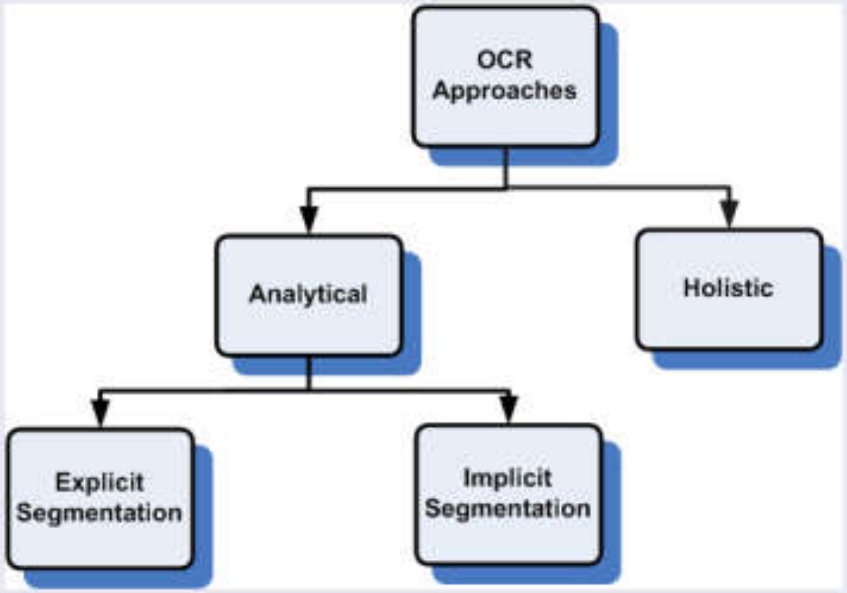
Touching diacritics and ligatures

Urdu Recognition Techniques

Recognition Techniques for Urdu Text

- Analytical or Segmentation-based
  - Recognition unit - Character
  - Explicit or Implicit
  - Explicit Segmentation: Challenging in cursive scripts
  - Implicit Segmentation: Huge training data
- Holistic or Segmentation-free
  - Recognition unit - Ligature

Taxonomy of Recognition Techniques



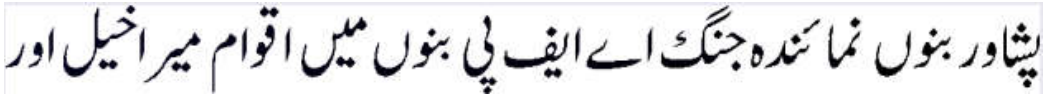
Benchmark Datasets

Urdu Printed Text Images (UPTI) dataset

- 10,063 text lines

Urdu Printed Text Images (UPTI) 2.0 dataset

- 1,20,000 text lines

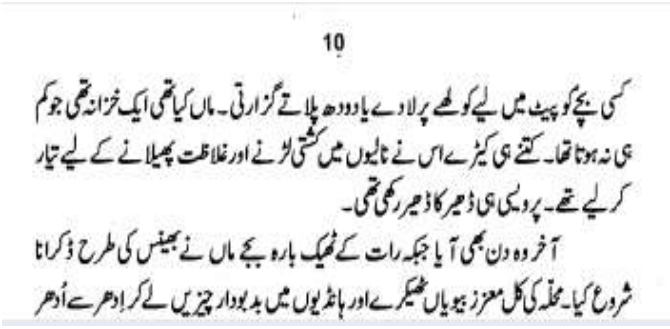


UPTI: Sample

CLE: SET-1 sample

Center for Language Engineering (CLE) dataset

- **SET-1:** 2,017 High Frequency Complete Ligatures (HFCLs) clusters
- **SET-2**: Challenging 2,912 images from 413 Urdu books



CLE: SET-2 sample

State of the Art-UPTI

Year	Authors	Technique	Features	Approach	Recognition Accuracy (%)
2013	Sabbour & Shafait et al.	Holistic	Deep Learning	BLSTM	91.00
2013	Adnan et al.	Analytical	Deep Learning	BLSTM	94.85
2016	Ahmed et al.	Analytical	Deep Learning	BLSTM	88.94
2016	Naz et al.	Analytical	Machine Learning	MDLSTM	94.00
2017	Naz et al.	Analytical	Machine Learning	MDLSTM	96.40
2016	Naz et al.	Analytical	Deep Learning based	MDLSTM	98.00
2017	Naz et al.	Analytical	Deep transfer Learning	LeNet-MDLSTM	98.12
2017	Ahmad et al.	Holistic	Deep Learning based	BLSTM	96.71
2017	Khattak et al.	Holistic	Machine Learning	HMM	92.26
2019	Akram & Hussain	Holistic	Machine Learning	HMM	98.37
2019	Khattak et al.	Holistic	Deep Learning	Scratched CNN	97.81

State of the Art-CLE

Year	Authors	Technique	Features	Approach	Recognition Accuracy (%)
2010	Javed et al.	Holistic	Machine Learning	HMM	92.00
2013	Javed & Hussain	Holistic	Machine Learning	HMM	92.73
2014	Hussain et al	Holistic	Machine Learning	HMM	97.87
2014	Akram et al.	Holistic	Machine Learning	HMM	86.15
2015	Hussain et al.	Analytical	Machine Learning	HMM	87.76
2015	Khattak et al.	Holistic	Machine Learning	HMM	97.93
2017	Akram et al.	Holistic	Machine Learning	Tesseract	97.87 97.71
2019	Akram & Hussain	Holistic	Machine Learning	HMM	95.58
2019	Khattak et al.	Holistic	Deep Learning	Scratched CNN	89.20

Commercial Urdu OCR

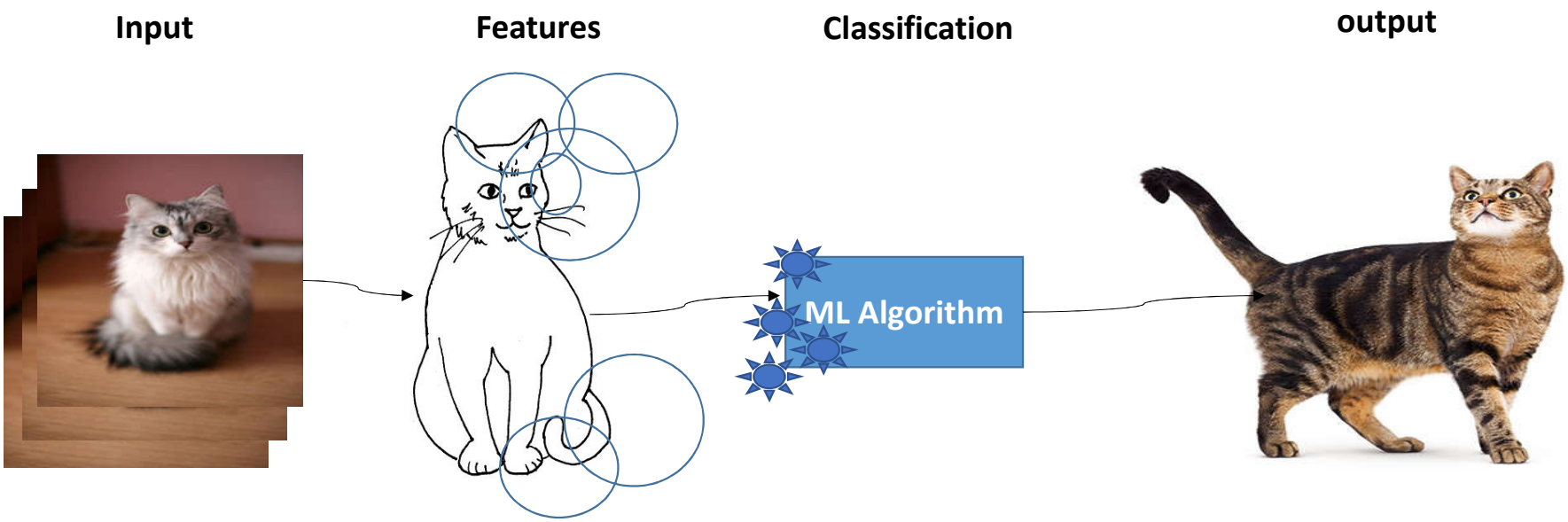
- Only available OCR
  - CLE Nastalique OCR 1.0.0[1] – Nasta'liq
  - i2OCR[2] (Free tool for 100 languages) – claim Urdu but Naskh

➢ **But still not satisfactory accuracy using unconstrained images of Nasta'liq text**

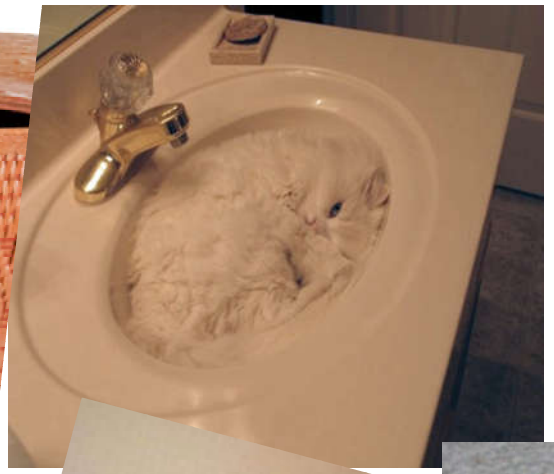
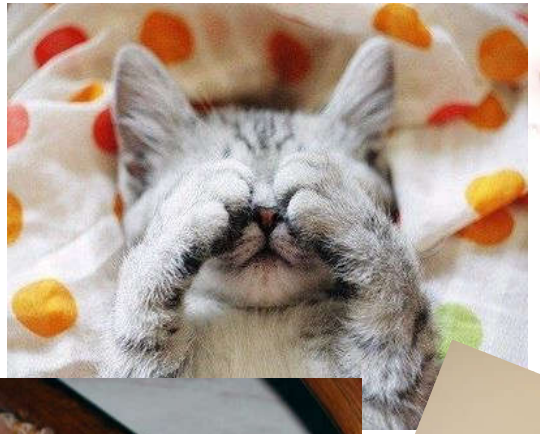
Naeem et al., Impact of Ligature Coverage on Training Practical Urdu OCR Systems, 14th IAPR International Conference on Document Analysis and Recognition, 2017.

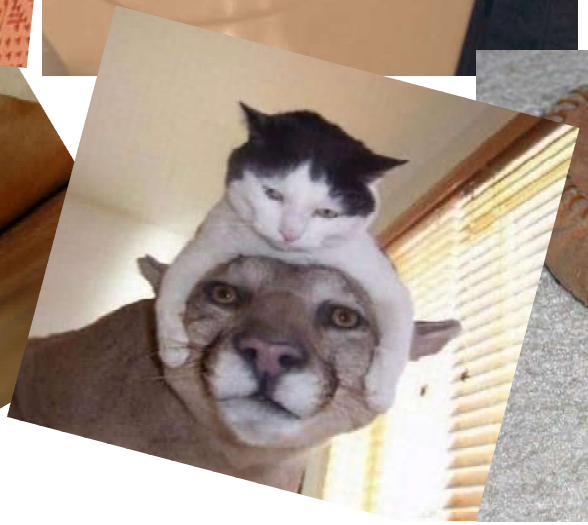
1. http://cle.org.pk/clestore/urduocr.htm
2. http://www.i2ocr.com/free-online-urdu-ocr

Traditional Machine Learning



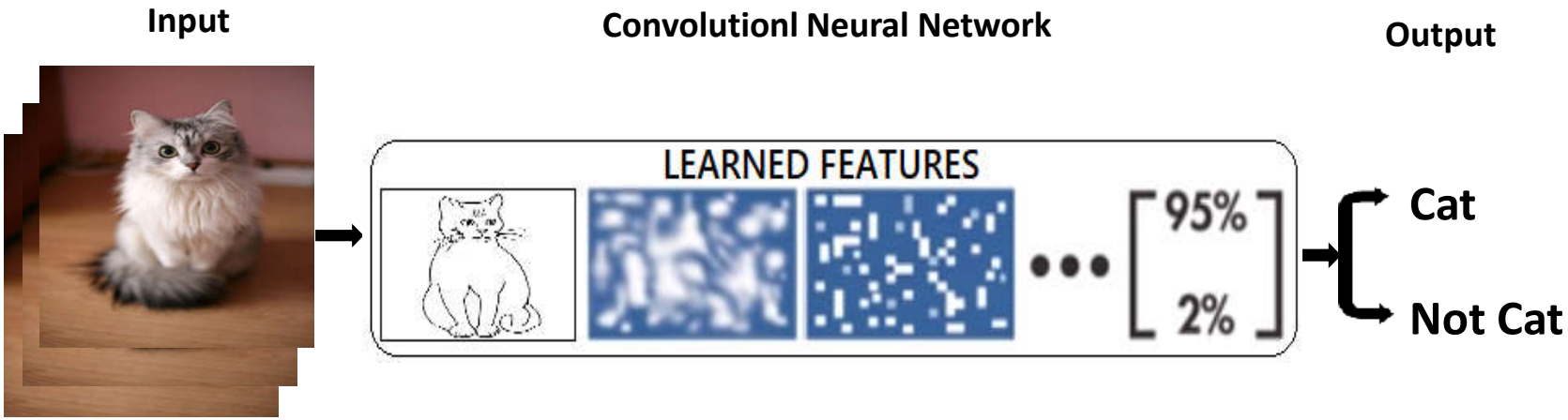
ML ➡ DL


DORA

Deep Learning



- Neuron: Atomic unit of network
- Network: a graph of neurons connected
  - DAG (most of networks)
  - Loops (some architectures)

Deep Learning

- Automatic features engineering from the raw input
- Handle very complex inputs and outputs
- Its end to end process
- Deep Learning architectures can be
  - CNN (Image)
    - AlexNet
    - GoogLeNet
    - VGGNet
    - ResNet
    - DenseNet
  - RNN (Sequence)
    - LSTM
    - BLSTM
    - MDLSTM
    - GRU

RNN

•Text of arbitrary length is a sequence of characters, and such problems are solved using RNNs and LSTM is a popular form of RNN.

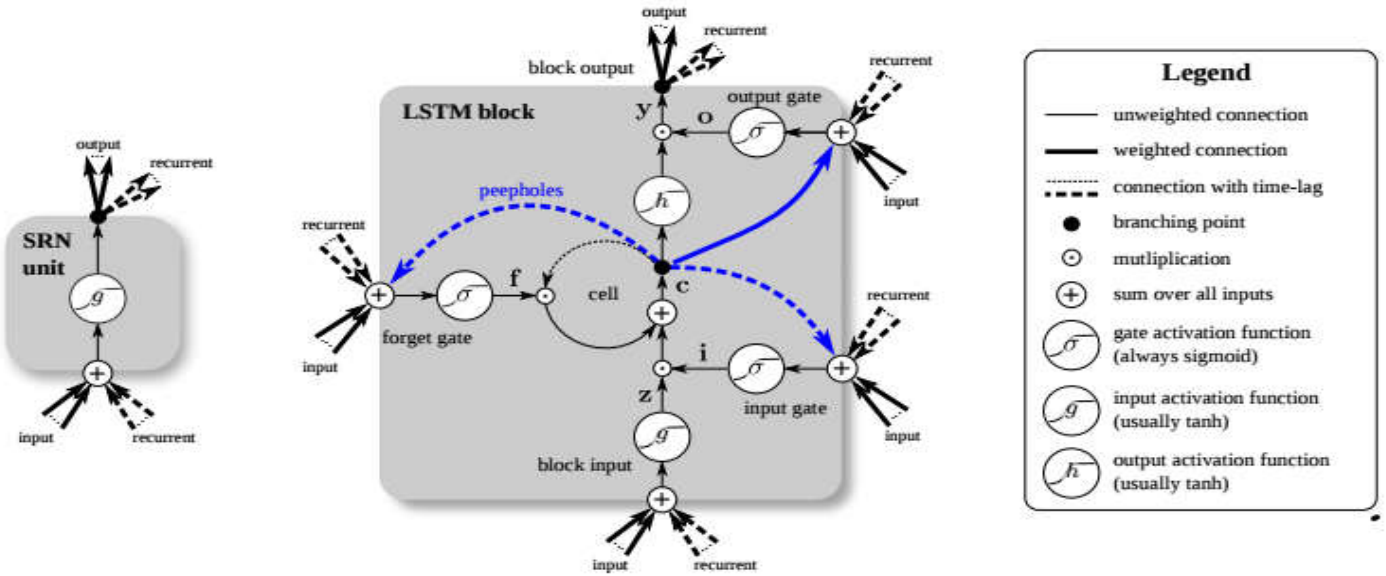


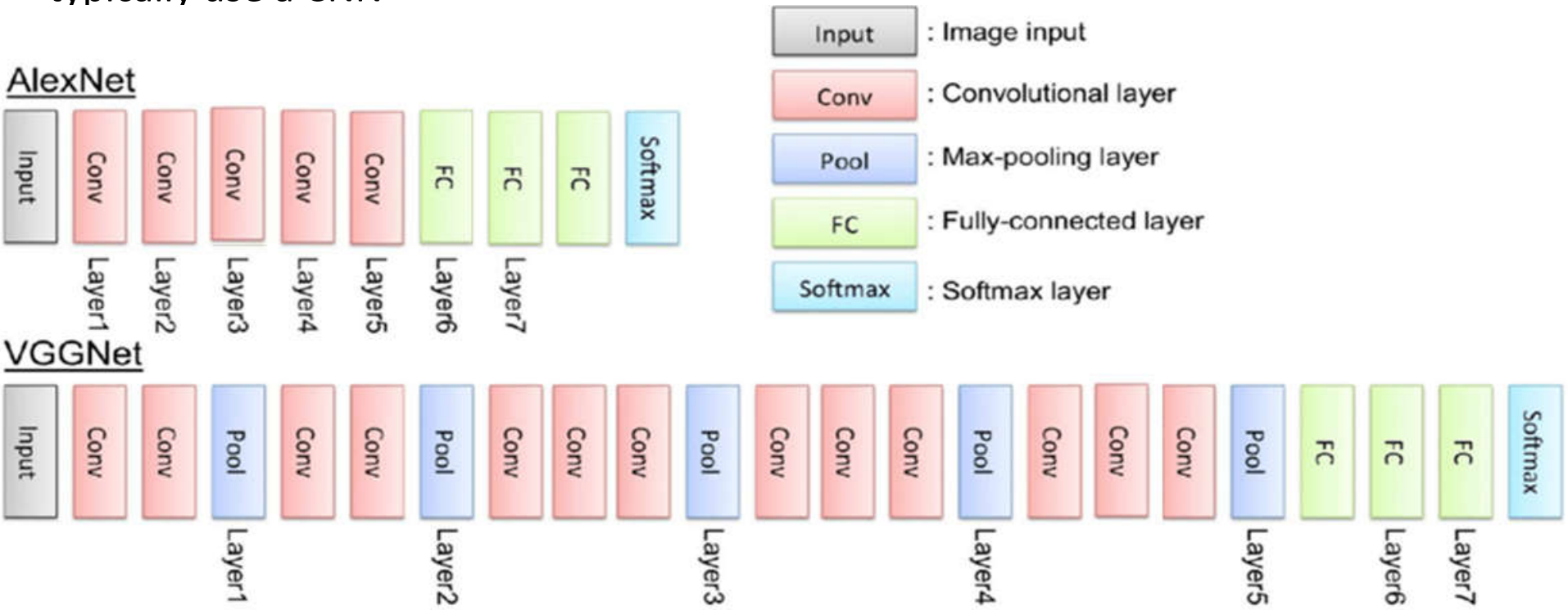
Figure 1. Detailed schematic of the Simple Recurrent Network (SRN) unit (left) and a Long Short-Term Memory block (right) as used in the hidden layers of a recurrent neural network.

Long Short Term Memory

- LSTM developed in 1997 - accuracy records in multiple applications domains
- Around 2007, LSTM based speech recognition, outperforming traditional models in certain speech applications
- In 2009, LSTM-CTC won contests of handwriting recognition
  - Chinese search engine Baidu
  - text-to-speech synthesis
  - Google Android
  - Google voice search
  - machine translation
- CNN-LSTM
  - automatic image captioning

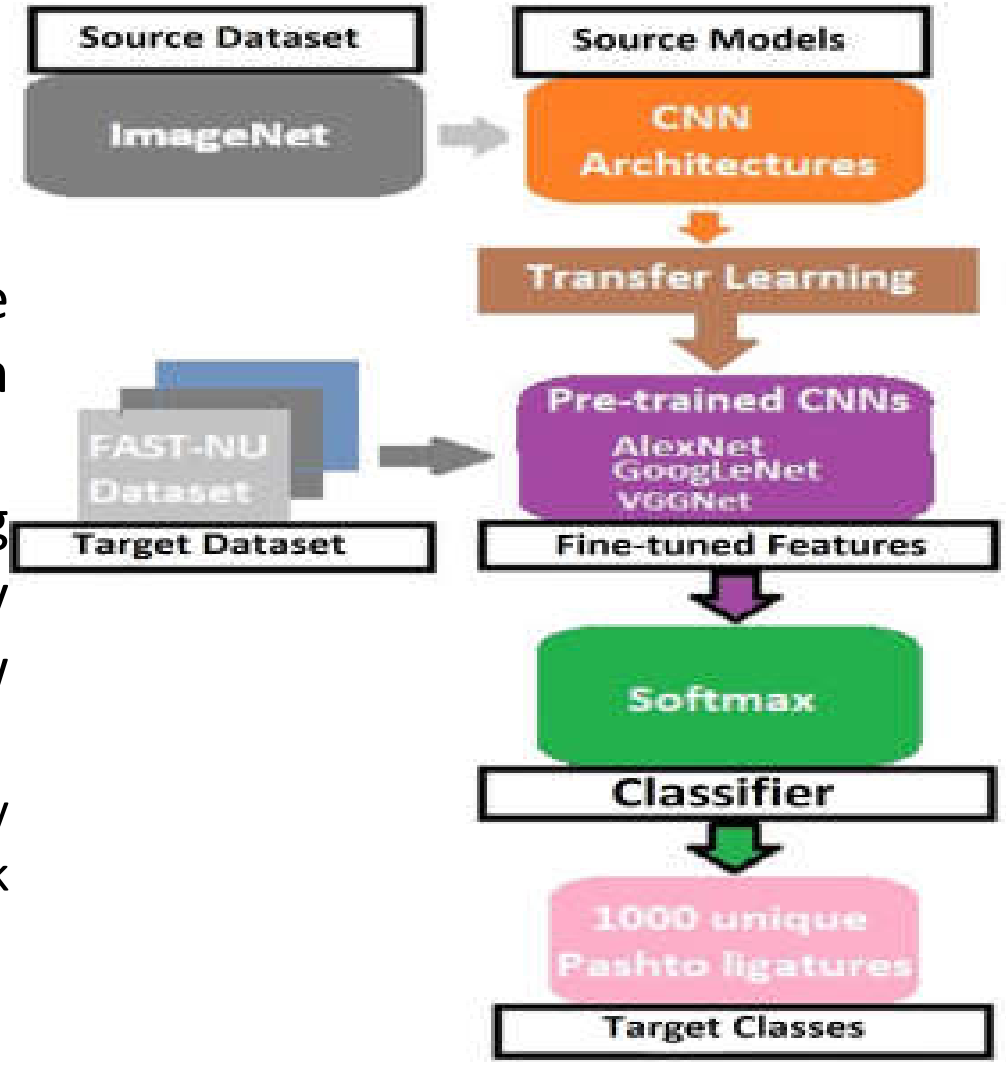
Convolutional Neural Network

- To recognize an image containing a single character, ligature, object or pattern, we typically use a CNN



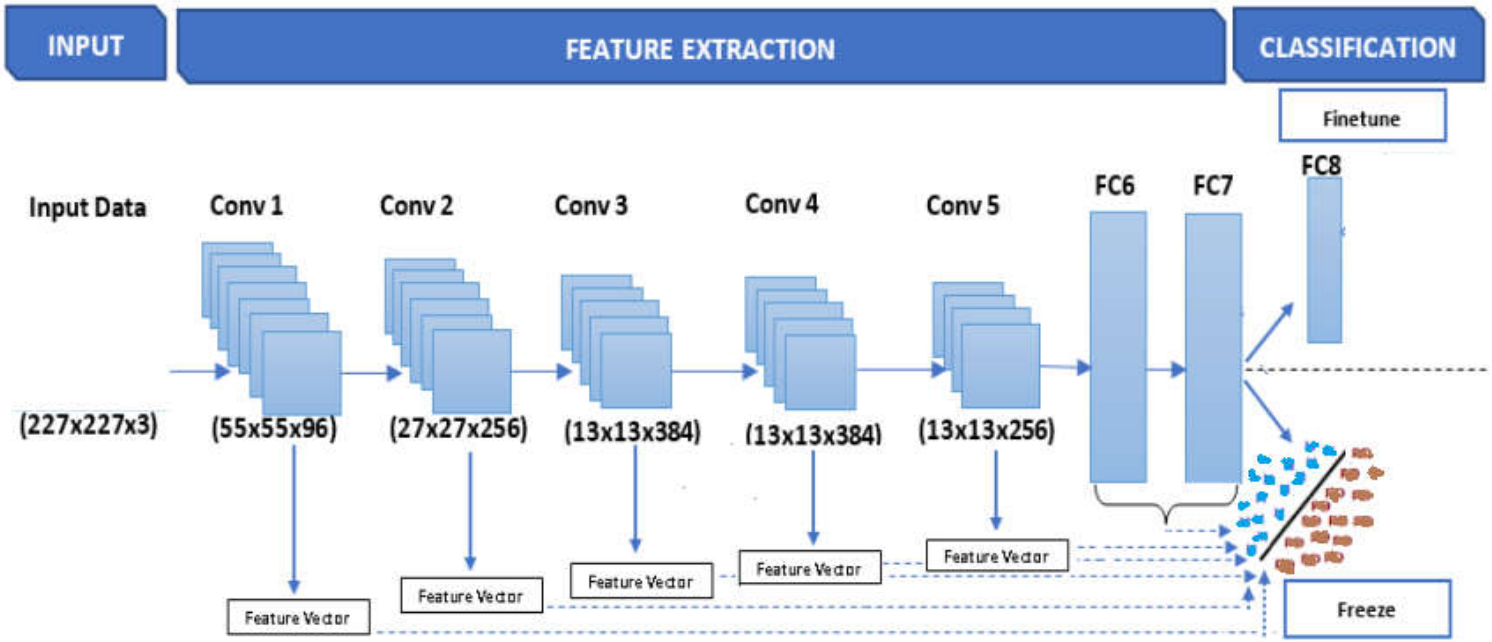
Transfer Learning

- TL an is approach in DL where knowledge is transferred from one model to anther.
- Re-using or transferring information from previously learned task for learning of new tasks
  - Has the potential to significantly improve the target task performance



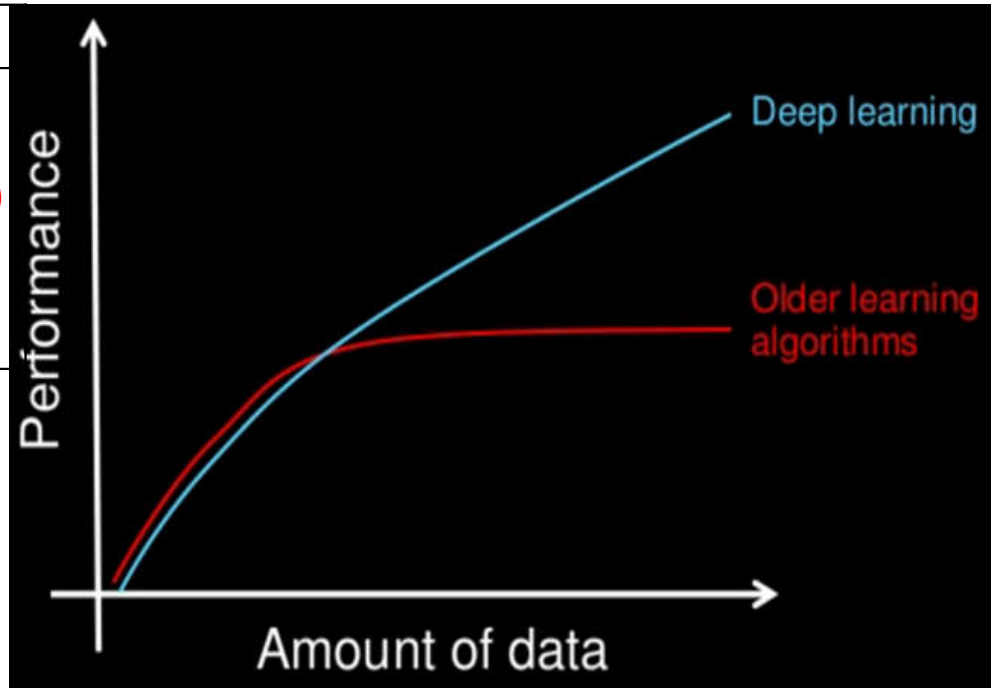
Freeze and Fine-tuned Features

- Use pre-trained weights of CNN architecture (usually ImageNet dataset) for classification of new task
- Fine-tuned the pre-trained weights of CNN architecture on target task



ML VS DL

ML	DL
Data (less) Parameters (less) Computational Power (less) Hardware (CPU) Execution time (less)	Data (large) Parameters (tens of million) Computational Power (high) Hardware (GPU) Execution time (more)



Deep Learning and Urdu OCR

- Ligature size is 26000 and training of 26K ligatures is nearly impossible with traditional ML
- Segmentation based approach very challenging due to laborious pre-processing and post-processing
- deep learning can overcome the issue
  - Labelling
  - Big data

G. S. Lehal, "Choice of recognizable units for urdu ocr," in *Proceeding of the workshop on document analysis and recognition*, 79–85, ACM (2012)

Dataset Labelling

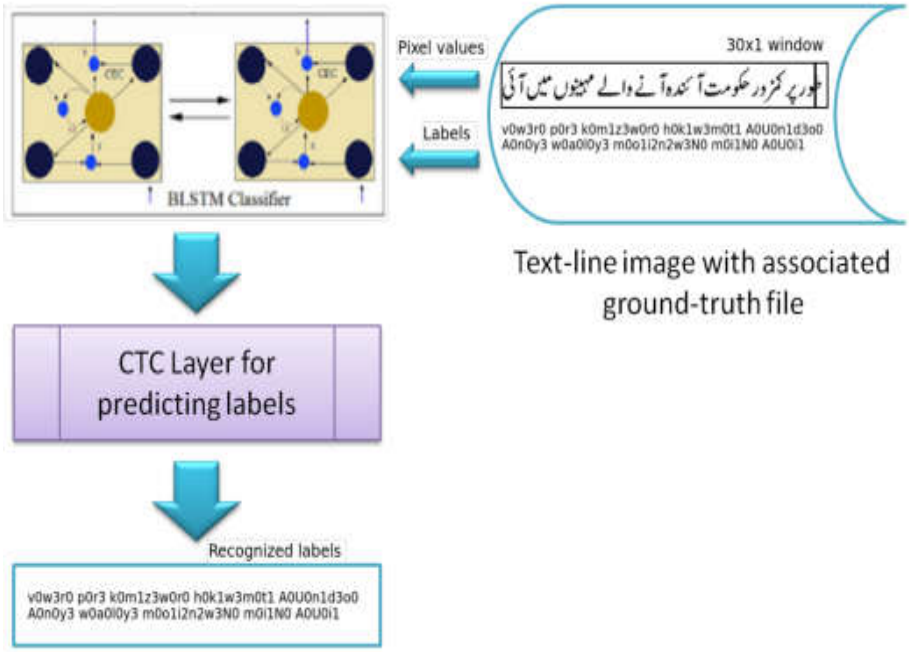
- Grouping and labeling of data according to Unicode
- Information extraction rate the same as on pure Unicode text

Basic characters		Labels	Unicode	Various shapes (glyph) of a basic character		Glyph with content in ligature
1	ا	alif	U+0627	Isolated	ا	
				Final	ـا	ـا_
2	ب	bay	U+0628	Isolated	ب	
				Initial	بـ	_بـ
				Medial	ـبـ	_بـ_
				Final	ـب	_ب

BLSTM-CTC based Urdu OCR

System 1:

- Automatic deep visional features extracted with 1 X 30 sliding window raw pixel values
- BLSTM network is fed with a 1D sequence of corresponding pixel values
- Connectionist Temporal Classification layer performs the output-transcription alignment
- Ave. accuracy is 94.85%

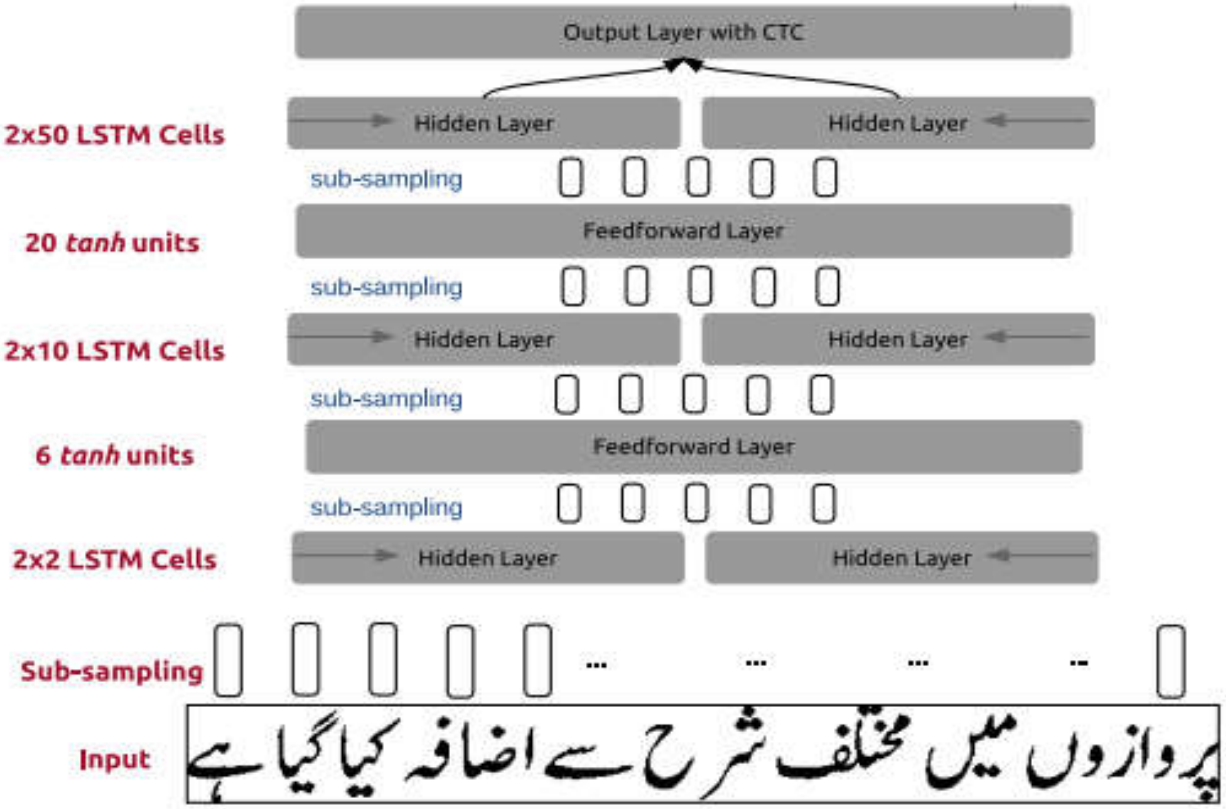


Ul-Hasan et al. "Offline printed Urdu Nastaleeq script recognition with bidirectional LSTM networks." In *2013 12th International Conference on Document Analysis and Recognition*, pp. 1061-1065. IEEE, 2013.

HSLSTM-CTC based Urdu OCR

System 2:

- The input from each layer is subsampled before feeding it to the next layer except the output layer
- Ave Char. Accuracy is 97.45%



Adnan ul hassan, "**Generic Text Recognition using Long Short-Term Memory Networks," PhD Thesis, 2016.**
"

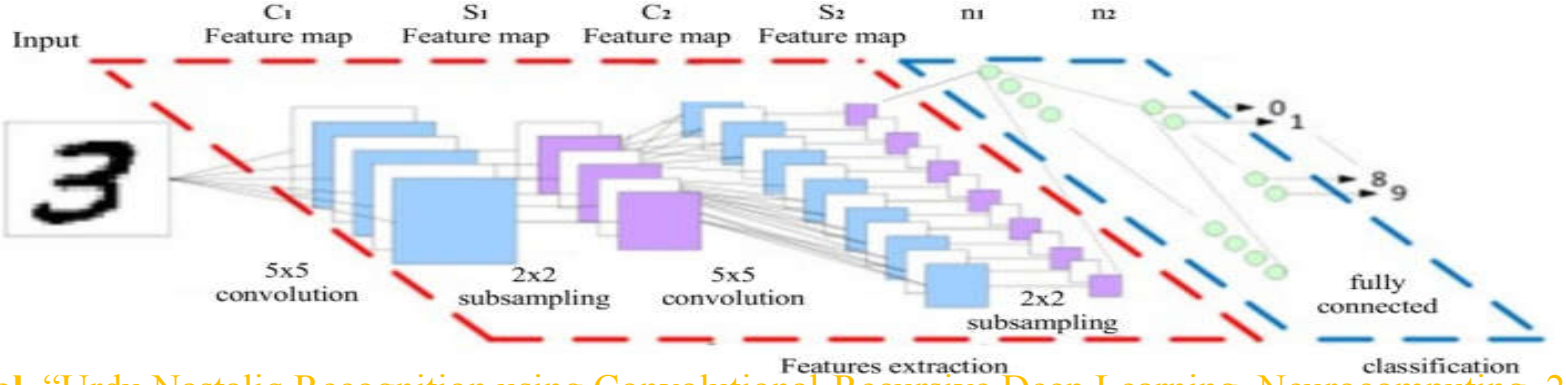
MDLSTM-CTC based Urdu OCR

- **System 3:**
  - 12 manual features extracted with 4 X 48 sliding window
  - MDLSTM input 1x1 info from each sliding window and contains 3 hidden layers of 248 nodes and 2 feed forward layers of 26 nodes
  - CTC transcribe unsegmented sequence data into 44 classes
  - Character level ave. accuracy is 94.97
- **System 4:**
  - Overlap sliding window
  - Character level ave. accuracy is 96.40

Naz et al. "Offline cursive Urdu-Nastaliq script recognition using multidimensional recurrent neural networks, Neurocomputing, vol. 177, pp. 228–241, 2016.

MDLSTM-CTC based Urdu OCR

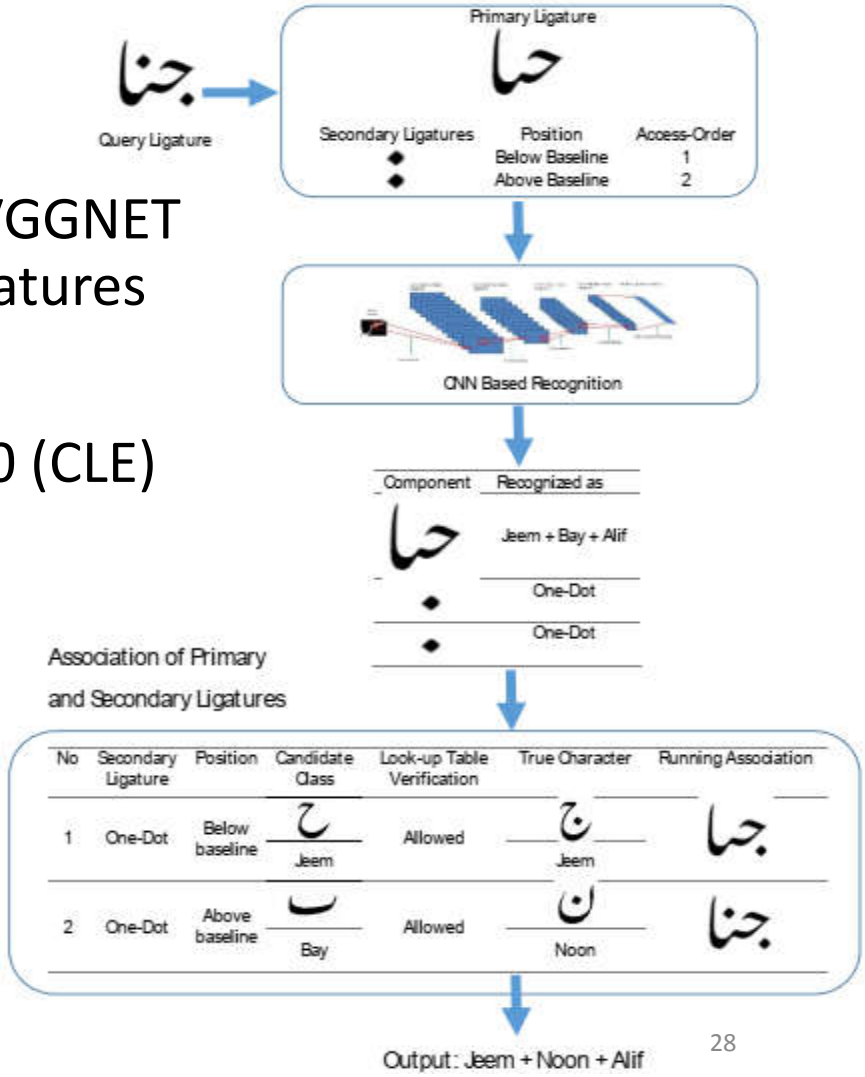
- **System 5:**
  - MDLSTM: automated features extracted from raw pixels
  - Ave. Test accuracy is 98%
- **System 6:**
  - Instead of statistical or automated MDLSTM based features, automatic features extracted from 1st conv layers of LeNet architectures and fed to MDLSTM-CTC
  - Ave test accuracy = 98.12%



Naz et al. "Urdu Nastaliq Recognition using Convolutional-Recursive Deep Learning ,Neurocomputing, 2017.

CNN based Urdu OCR

- Transfer Learning using AlexNet and VGGNET (16 & 19) using Freeze and Fine-tuned Features
- Investigate ConvNet from scratched
- Ave. highest Accuracy 94.78 (UPTI), 88.10 (CLE)



Uldin et al.,"Recognition of printed Urdu ligatures using convolutional neural networks," *Journal of Electronic Imaging* 28(3), 2019

Table 5 Recognition rates on CLE and UPTI datasets

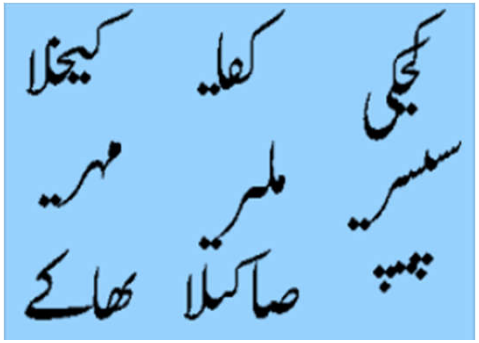
Network	Mode	CLE (HFL)	CLE (Books)	UPTI	CLE (Books) + UPTI	
					Query:CLE	Query:UPTI
Alexnet	Pre-trained	94.39%	87.10%	92.08%	86.70%	89.28%
	Fine tuned	95.78%	88.70%	94.88%	87.60%	91.81%
VGG16	Pre-trained	93.25%	83.50%	90.84%	82.50%	86.73%
	Fine tuned	94.10%	85.40%	93.33%	84.40%	87.26%
VGG19	Pre-trained	92.51%	82.60%	91.58%	82.60%	88.60%
	Fine-tuned	93.90%	84.70%	94.22%	83.90%	89.96%
Proposed	-	**98.30%**	**89.20%**	**97.81%**	**88.10%**	**94.78%**

Uldin et al.,"Recognition of printed Urdu ligatures using convolutional neural networks," *Journal of Electronic Imaging* 28(3), 2019

Actual-Predicted

Query	Mismatched	Query	Mismatched
سلیم	سلم	سر	سر
عبا	عبا	سہے	سہے
سل	سل	لسد	لسد
لس	لس	ھے	ھے

Ligatures having Visually Similar Shapes Resulting in a Mismatch



False joining of secondary ligatures with the respective or neighboring primary ligature

False re-association of dots/diacritics with the primary ligatures

Query Complete Ligature	Falsely Formed Complete Ligature
عظم	غظم
عظیم	غظیم
جعفر	جغفر
تحفظ	تخفظ
بعض	بغض

Future of OCR

- Future of OCR is deep learning
- To work on for end to end system for Urdu text
  - Open Source Engines/tools
    - Tesseract
    - Ocropy 2.0
    - Umaru
    - Calamari
    - Attention-OCR
  - At word or ligature level – CRNN or CNN + post processing
- The error rates can be reduced further using language model

Tesseract

- An Open Source OCR Engine for multiple scripts and languages
  - https://github.com/tesseract-ocr/tesseract
- In 2005, it was originally developed at HP-Lab, Since 2006 it is developed by Google
- Tesseract 3.x is based on traditional computer vision algorithms
  - The Tesseract classifier has adapted for complex script (Urdu Nastalique, Acc. 97.87)
- Tesseract 4.x based on DL neural net (LSTM) based OCR engine which is focused on text line recognition
  - Need to explore for Urdu text

Ain et. al, "Adapting Tesseract for Complex Scripts: An Example for Urdu Nastalique,"

Future of OCR (cont…)

- To work on for end to end system for Urdu text
  - Open Source Engines/tools
    - Tesseract
    - Umaru
    - ocropy
    - Calamari
    - Attention-OCR
  - At word or ligature level – CRNN or CNN + post processing
- The error rates can be reduced further using language model

Umara

- An OCR-system based on torch using the technique of LSTM/GRU-RNN, CTC and referred to the works of rnnlib and clstm
  - https://github.com/edward-zhu/umaru
- RNNLIB is a recurrent neural network library for sequence learning problems and it has proven particularly effective for speech and handwriting recognition
  - https://sourceforge.net/p/rnnl/wiki/Home/
- CLSTM is small C++ implementation of LSTM networks, focused on OCR
  - https://github.com/tmbdev/clstm

A Graves et al, “Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks”, ICML, 2006

Future of OCR (cont…)

- To work on for end to end system for Urdu text
  - Open Source Engines/tools
    - Tesseract
    - Umaru
    - Ocropy
    - Calamari
    - Attention-OCR
  - At word or ligature level – CRNN or CNN + post processing
- The error rates can be reduced further using language model

Ocropy

- **OCRopus** – ML-based tools for document analysis and OCR in python, called Ocropy
- LSTMs based its engine is referred as Ocropy2
  - https://github.com/tmbdev/ocropy
  - https://github.com/tmbdev/ocropy2
- OCRopus is a collection of document analysis programs, not a turn-key OCR system
  - Tools for ground truth labeling
  - Pluggable layout analysis
  - Pluggable character recognition (without language modeling, separate command for binarization and recognition)
  - Pluggable language modeling
  - Unicode and ligature support
- training data should be in .png/.gt.txt files

Developer: Thomas Breuel, DFK

Future of OCR (cont…)

- To work on for end to end system for Urdu text
  - Open Source Engines/tools
    - Tesseract
    - Umaru
    - Ocropy
    - Clamari, SynthText . . .
  - At word or ligature level – CRNN or CNN + post processing
    - Attention-OCR
- The error rates can be reduced further using language model

https://github.com/da03/Attention-OCR
https://github.com/ankush-me/SynthText
C Wick et al, "Calamari - A High-Performance Tensorflow-based Deep Learning Package for Optical Character Recognition", 2018.

Need Attention for Urdu text

- Wild/Scene text recognition
- Camera based character recognition
- Online handwritten recognition
- Recognition system for video (Urdu) text
- To expose the recognition system to the more challenging problem of Urdu text like newspapers
- There is also need to focus on document imaging for book, specially old book, that will have a lot challenges to image acquisition

First need to develop dataset

Thank You

Questions /Answers