Deep learning based OCR engine for the Indus script

Satish Palaniappan
(IMSc & Qube Cinemas)

Where it all started?

Computational Epigraphy

Epigraphy is the study of ancient inscriptions, and the place where mathematics and computer science concepts meet epigraphy is called computational epigraphy.

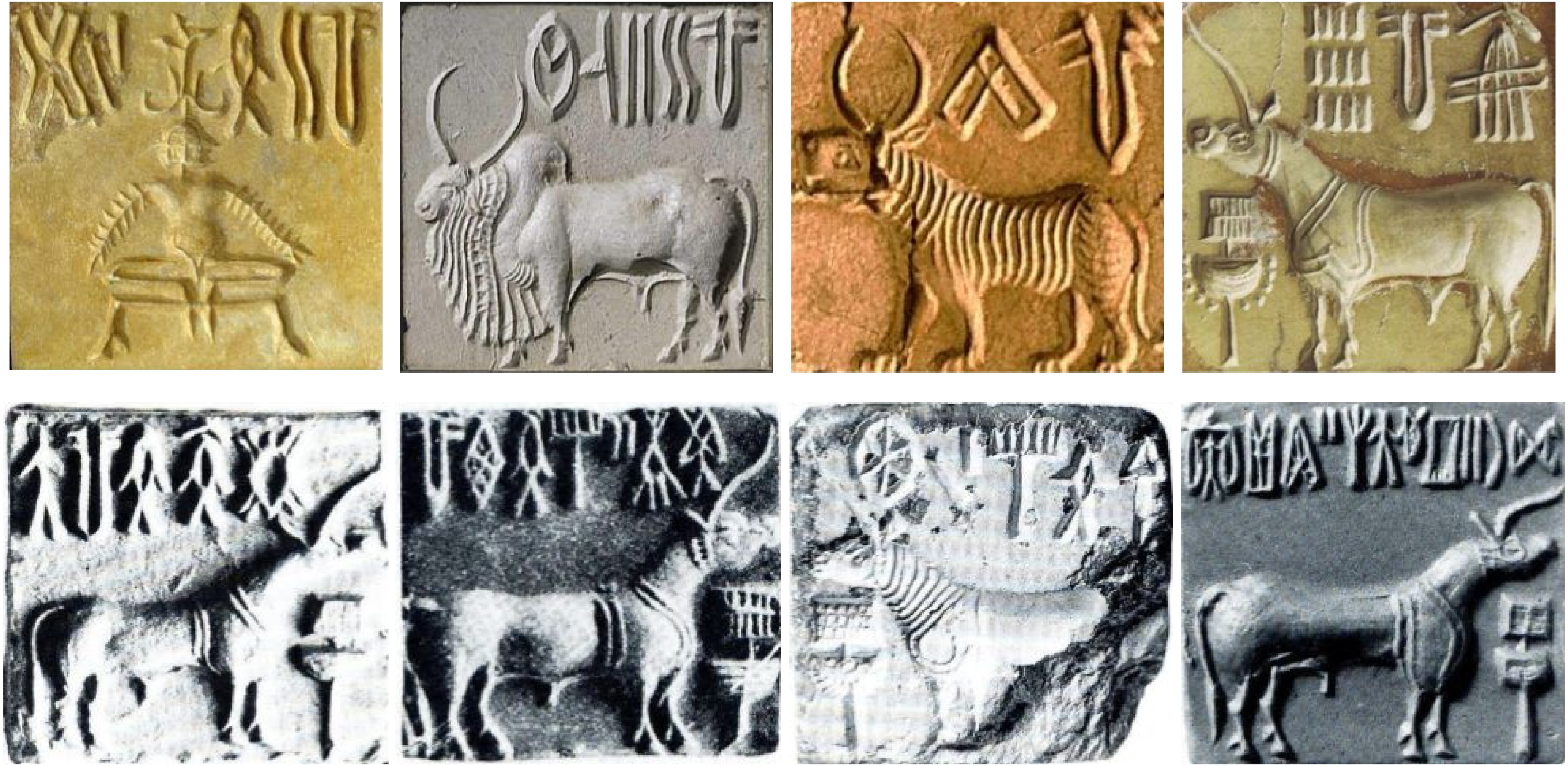
The Indus Script

- Indus valley / Harappan civilization
- Largest and one of the most ancient civilizations known to mankind
- Major Indus valley sites: (Northwestern regions of India)
    - Harappa, Mohenjo-Daro
    - Chanhu-Daro, Lothal, Kalibangan …
- Inscription’s form factor:
    - Stamp seals and Sealings
    - Amulets, stone tablets, pottery …
- The script:
    - Around 3700 text inscriptions according to the M77 corpus, with an average of 5 graphemes per text

Why Still Undeciphered?

- Paucity of long texts (rarely with 14 graphemes per text)
- Absence of parallel or bilingual text
- No definite knowledge about the underlying language
- The number of graphemes and the very less data (417 symbols with just 3700 texts documented)

Sample Indus Seals



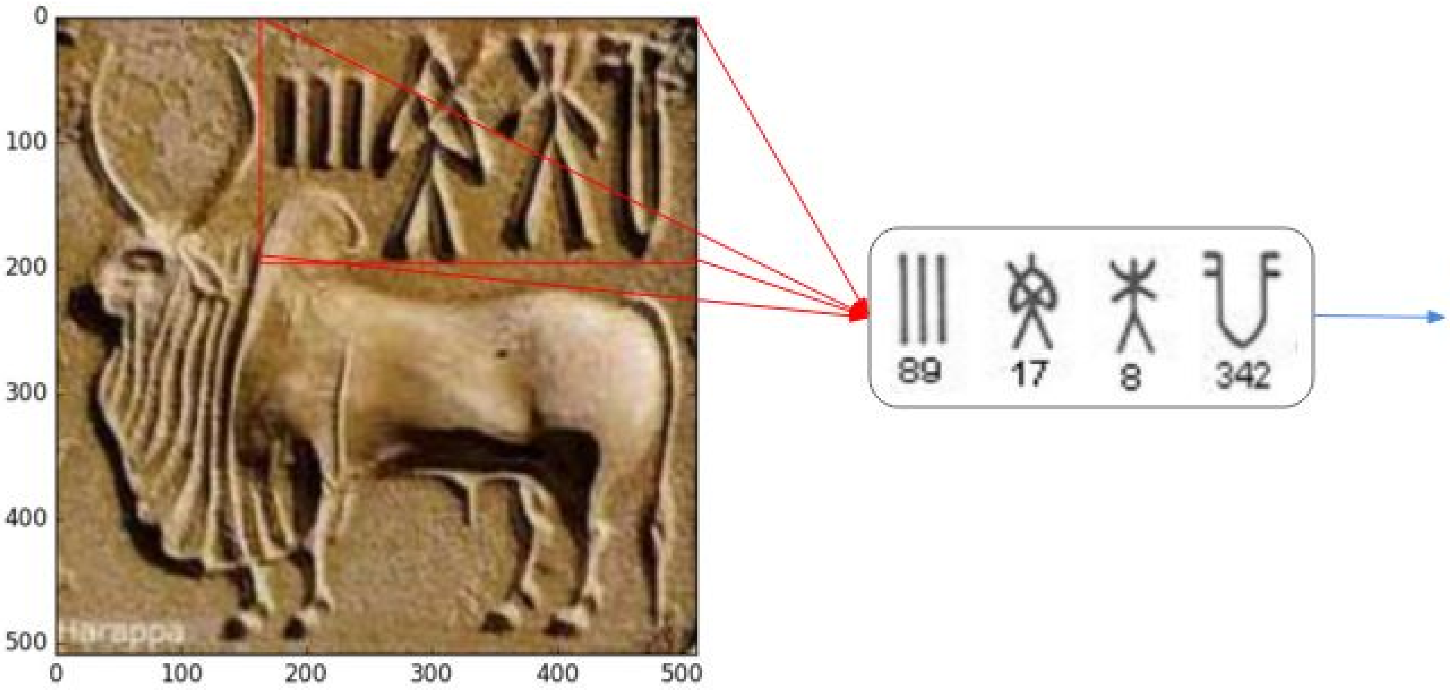
Machine Learning in the Study of Indus Script

- In the past, ML has been used for:
  - Classification
    - Based on patterns
    - Based on graphemes
  - Graphemic pattern search
  - Linguistic structure
  - Markov models

Corpus Formulation

- Need
  - Bottleneck to all ML based research
  - We have been using a 40 years old dataset, with no recent updates
  - Indus script will remain an enigma
- Challenges
  - Extremely laborious human (expert) effort
  - Time consuming to standardize for use
  - Other political issues

Problem Overview

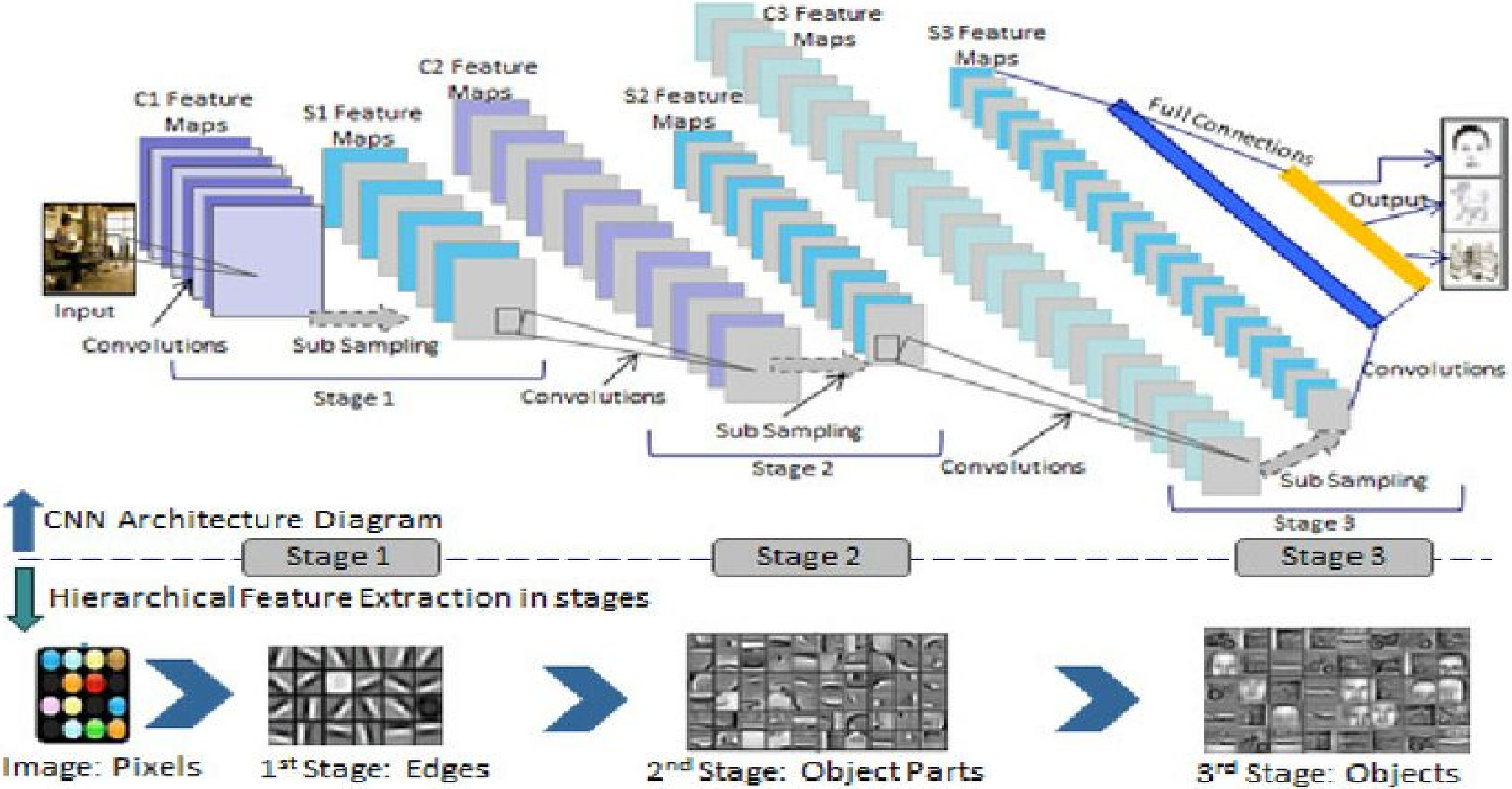


Output:
89, 17, 8, 342

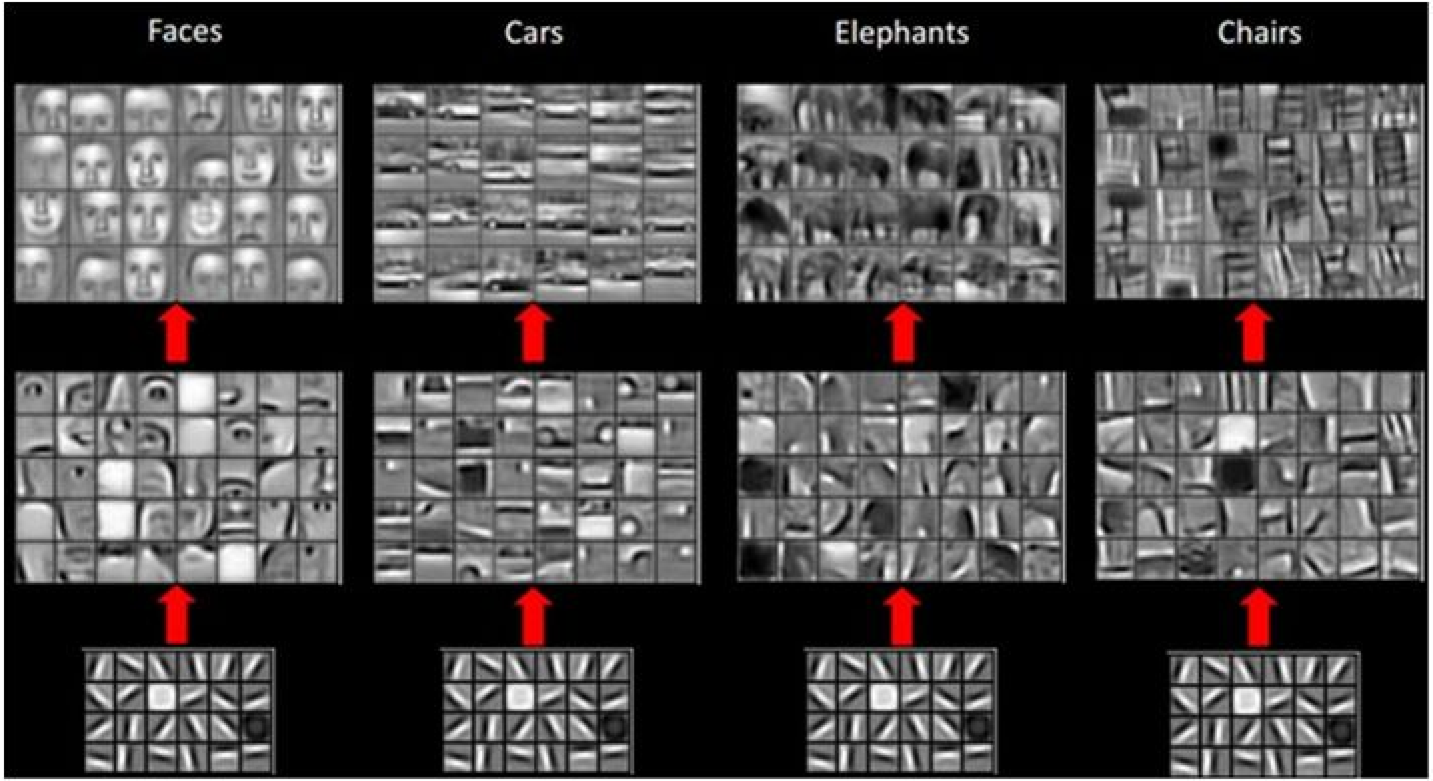
Why Deep Learning?

- Complexities in building an OCR engine for reading the Indus script
    - Wear and tear - The artifacts are nearly 4000 year old relics
    - Form factor of the artifacts
    - Very less data with more characters (symbols) to recognize
    - No fixed character set
    - Minute differences between the symbols, account for a completely different representation
    - More undocumented symbols

Handcrafting the features = Nightmare!



Transfer Learning and Fine Tuning

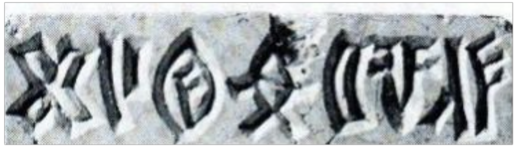


Data Augmentation

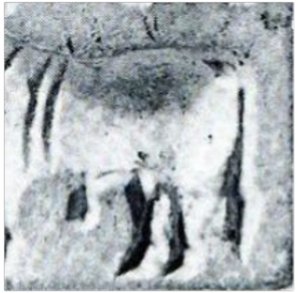
- To compensate for the meager data, we devised certain image augmentation techniques in addition to what Keras has.
  - Vertical and horizontal flips
  - Shear, Crop, Swirl
  - Rotate, Scale, Translate
  - Randomised artificial lighting

Datasets Used

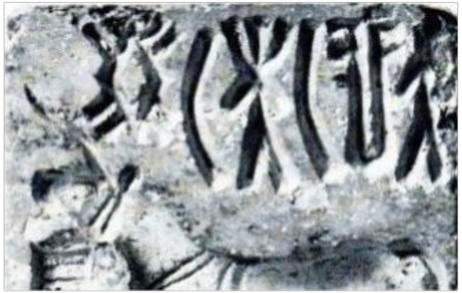
Text-NoText Dataset



Text

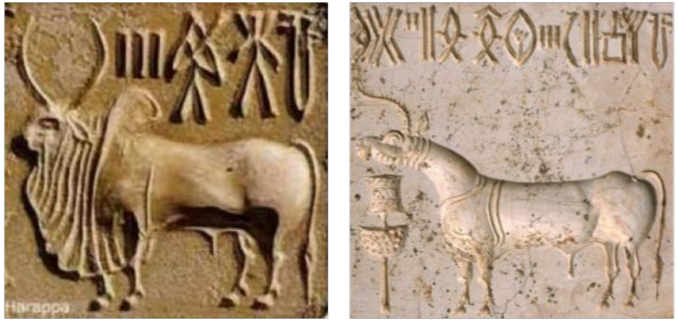


No-Text

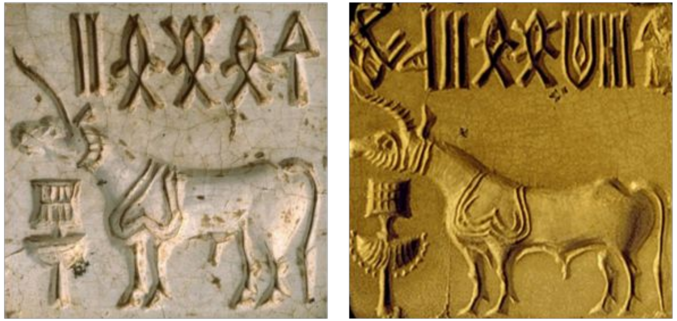


Both

Symbols Dataset



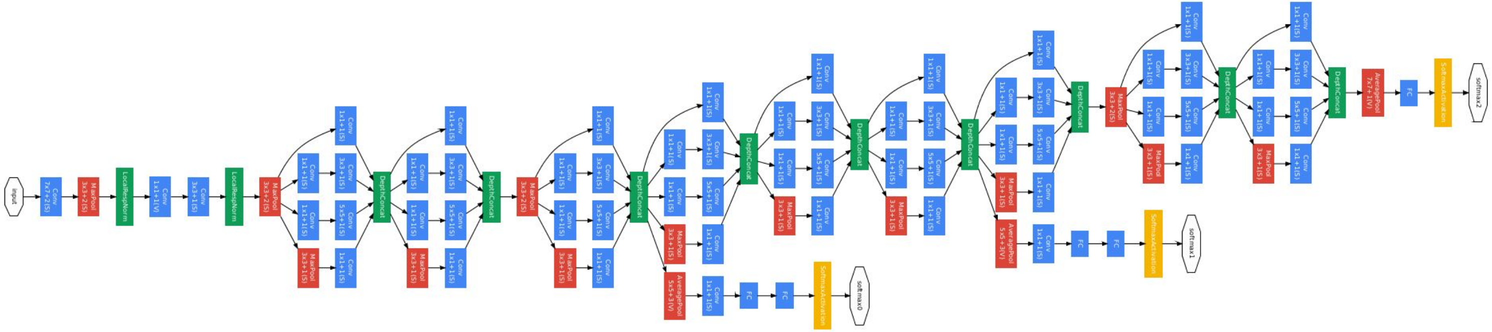
'Jar' sign present



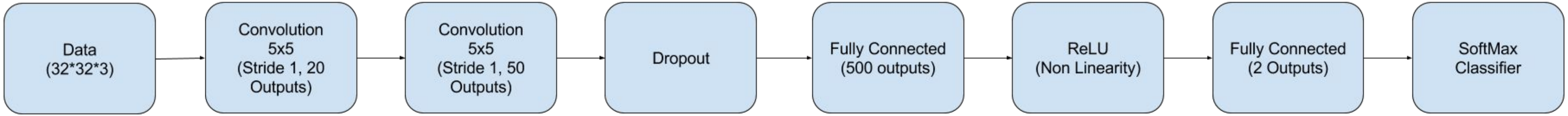
'Jar' sign absent

Base CNN Architectures Used

GoogLeNet



SymbolNet



Data (32*32*3)

Convolution 5x5 (Stride 1, 20 Outputs)

Convolution 5x5 (Stride 1, 50 Outputs)

Dropout

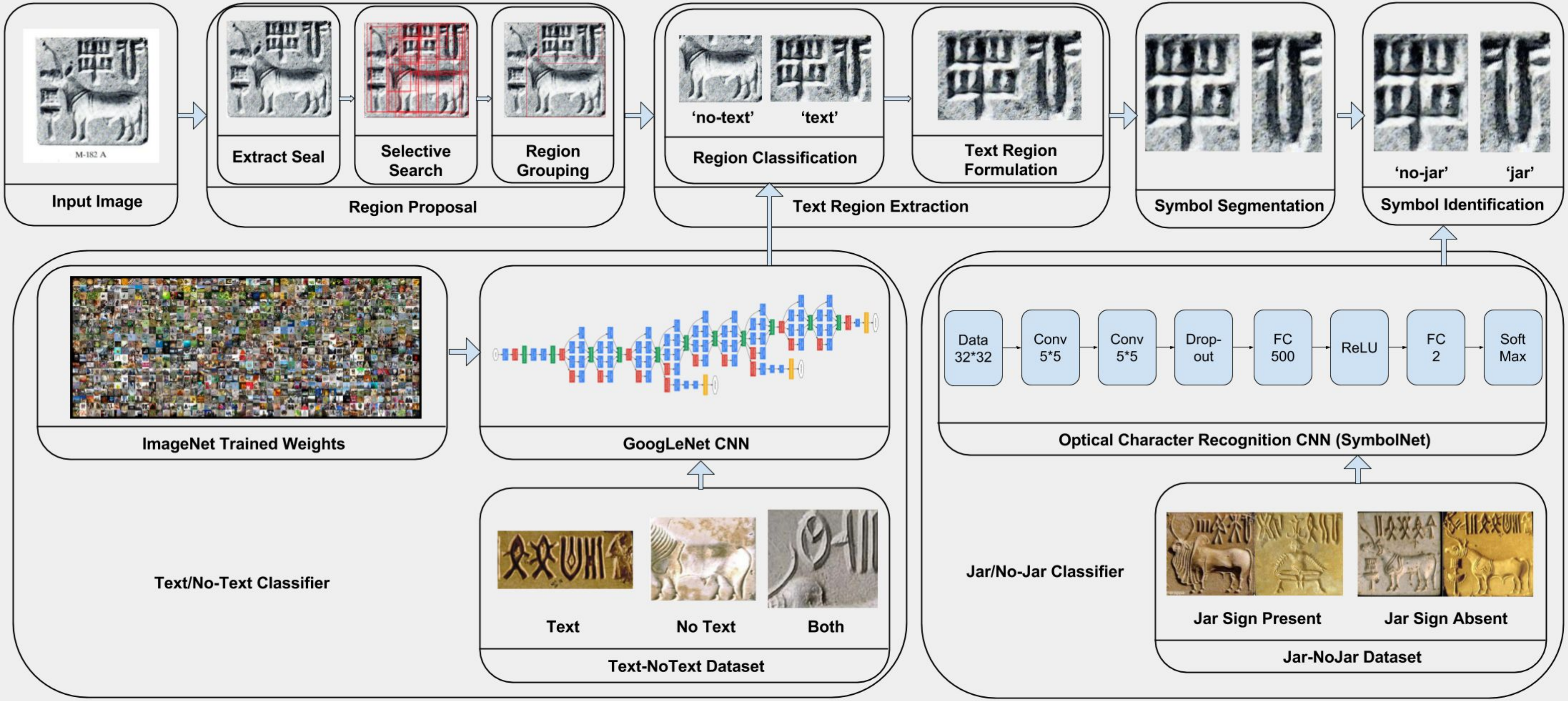
Fully Connected (500 outputs)

ReLU (Non Linearity)

Fully Connected (2 Outputs)

SoftMax Classifier

The Pipeline



M-182 A

Input Image

Extract Seal

Selective Search

Region Grouping

Region Proposal

‘no-text’

‘text’

Region Classification

Text Region Formulation

Text Region Extraction

Symbol Segmentation

‘no-jar’

‘jar’

Symbol Identification

ImageNet Trained Weights

GoogLeNet CNN

Text/No-Text Classifier

Text

No Text

Both

Text-NoText Dataset

Data 32*32

Conv 5*5

Conv 5*5

Drop-out

FC 500

ReLU

FC 2

Soft Max

Optical Character Recognition CNN (SymbolNet)

Jar/No-Jar Classifier

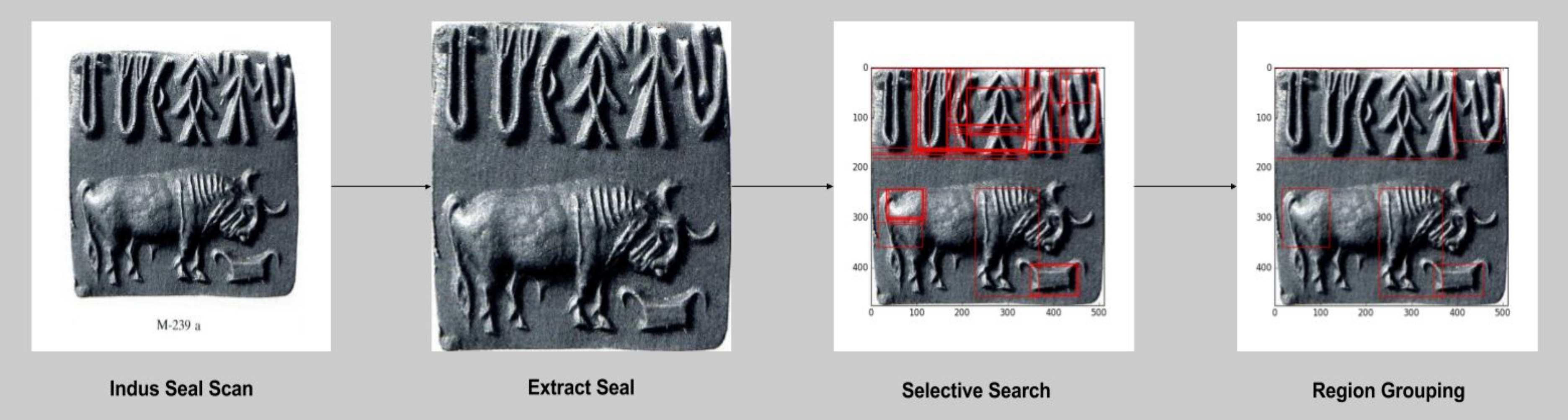
Jar Sign Present

Jar Sign Absent

Jar-NoJar Dataset

Region Proposal

Proposes regions that have a high probability of containing a symbol, animal, deity, or any iconographic element.



Indus Seal Scan

Extract Seal

Selective Search

Region Grouping

Extract Seal

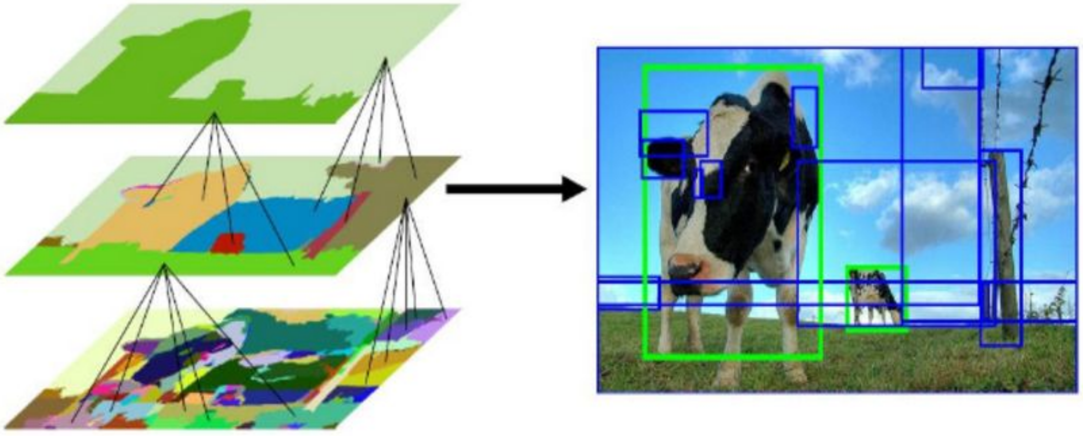
Removes the irrelevant background information from the input artifact image.

- Steps
    - Grayscaled and smoothened using multi-scale gaussian filter
    - Threshold the image at the mean pixel value of the background
    - Optimized canny edge detection

Selective Search

Proposes an array of all possible regions, likely to hold Indus script symbols, depictions of animals, etc.

- Combines the advantages of exhaustive search and segmentation
- Hierarchical grouping of the region proposals based on color, texture, size and fills
- Grid search over the four Selective Search parameters to use the Scale, Min Size, Min Area, and Sigma



Region Grouping

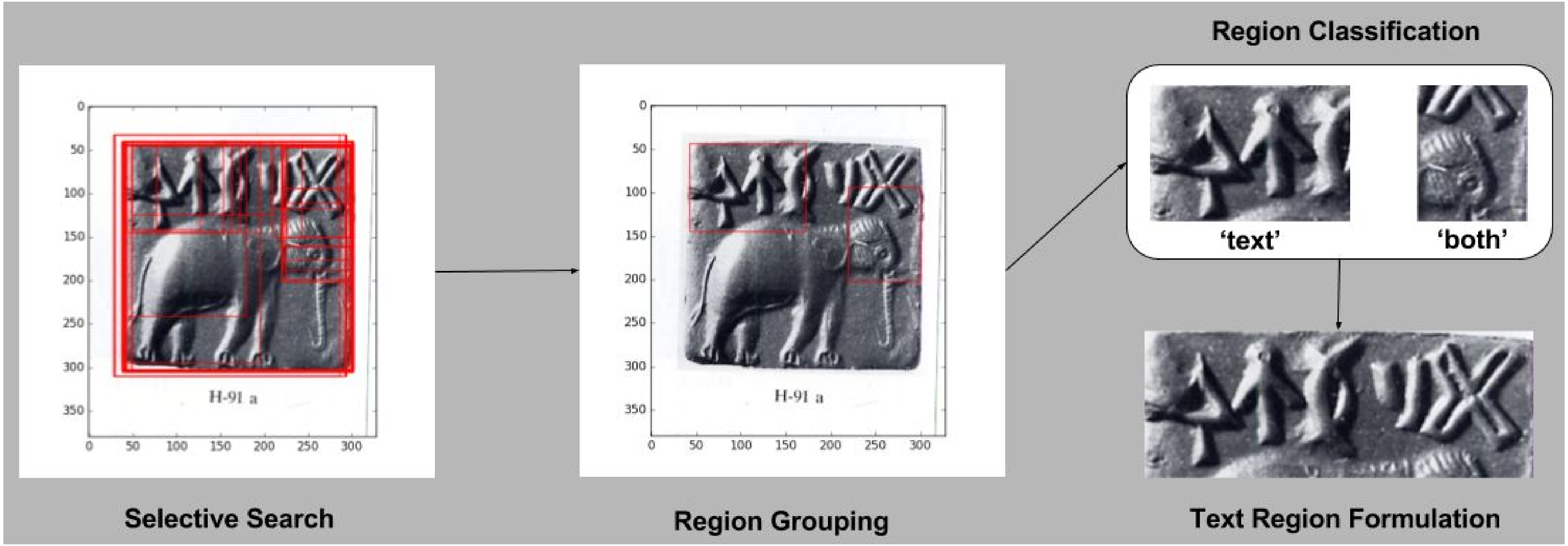
Improvises the quality of the regions proposed by Selective Search (four-level region grouping and filtering hierarchy)

- Merge concentric proposals
- Contained boxes removal
- Draw super box
- Draw extended super box

The last two levels leverage the prior information that the text regions are contiguous, being mostly arranged along a same line or axis rather than randomly distributed in space.

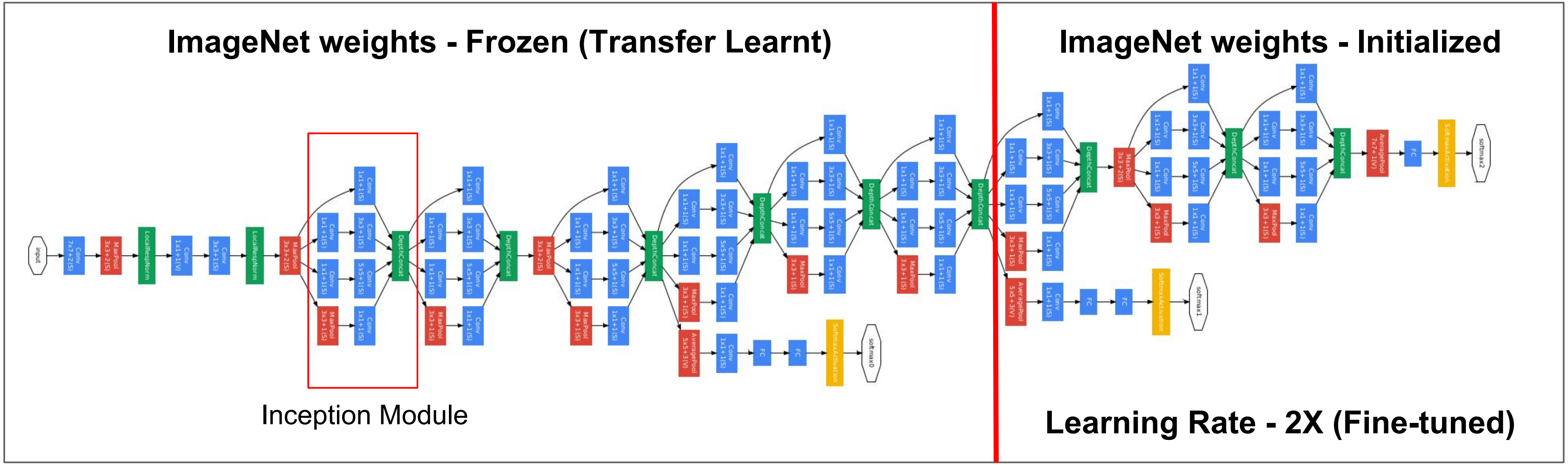
Text Region Extraction

Produces exact text-only regions by eliminating the non-graphemic parts off of the region proposals in hand.



Region Classification

All the regions are classified into types: “Text”, “No-Text” or “Both”.



ImageNet weights - Frozen (Transfer Learnt)

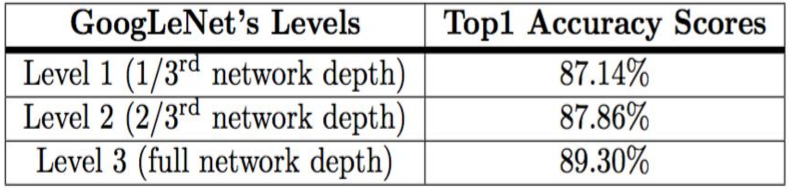
ImageNet weights - Initialized

Inception Module

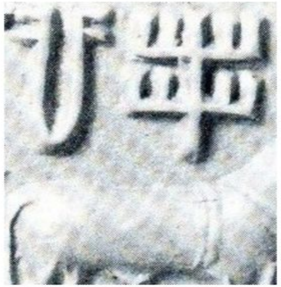
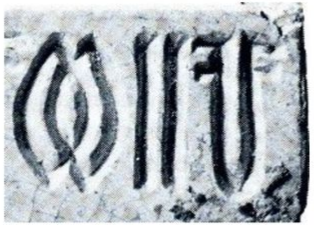
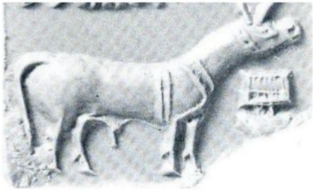
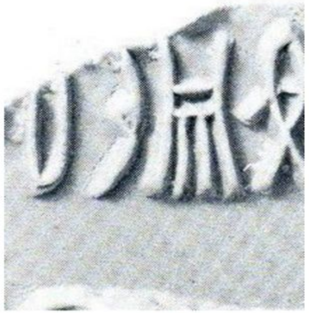
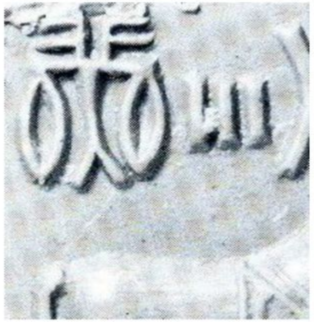
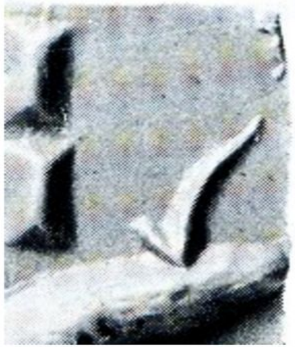
Learning Rate - 2X (Fine-tuned)

The Region Classification CNN's Results

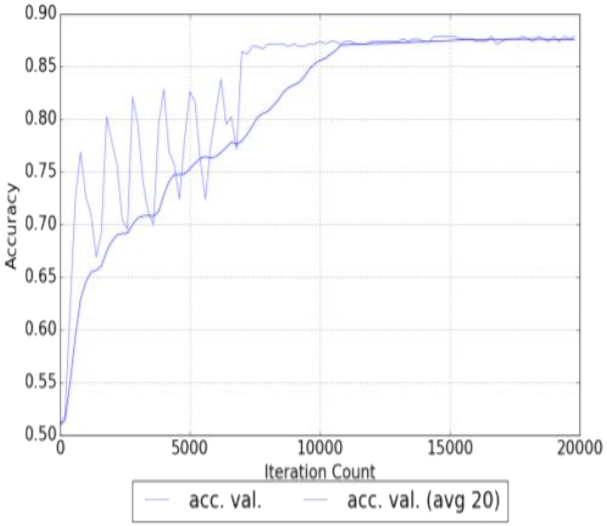
Top-1 Accuracy Scores



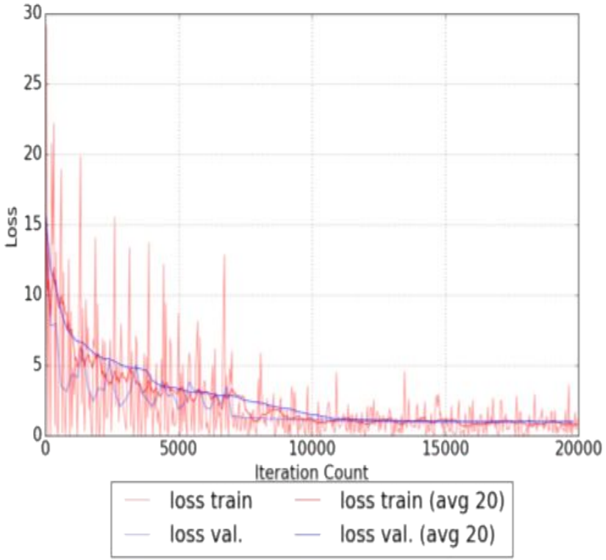
GoogLeNet's Levels	Top1 Accuracy Scores
Level 1 (1/3rd network depth)	87.14%
Level 2 (2/3rd network depth)	87.86%
Level 3 (full network depth)	89.30%

		
Expected Label: Both **Actual Label:** Both	**Expected Label:** Text **Actual Label:** Text	**Expected Label:** No Text **Actual Label:** No Text
		
Expected Label: Text **Actual Label:** Text	**Expected Label:** Both **Actual Label:** Text	**Expected Label:** Both **Actual Label:** No Text

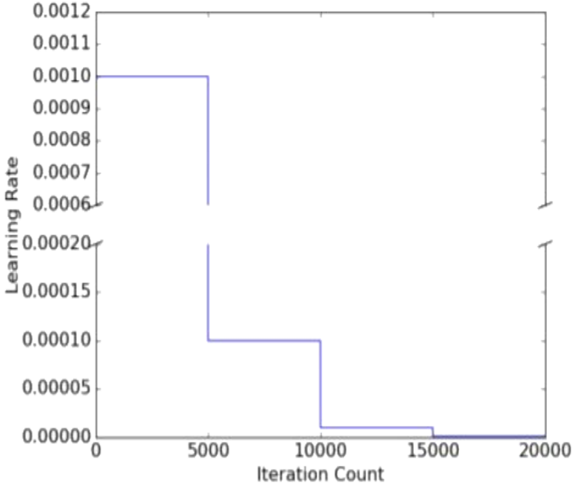
The Region Classification CNN's Graphs



(a)



(b)



(c)

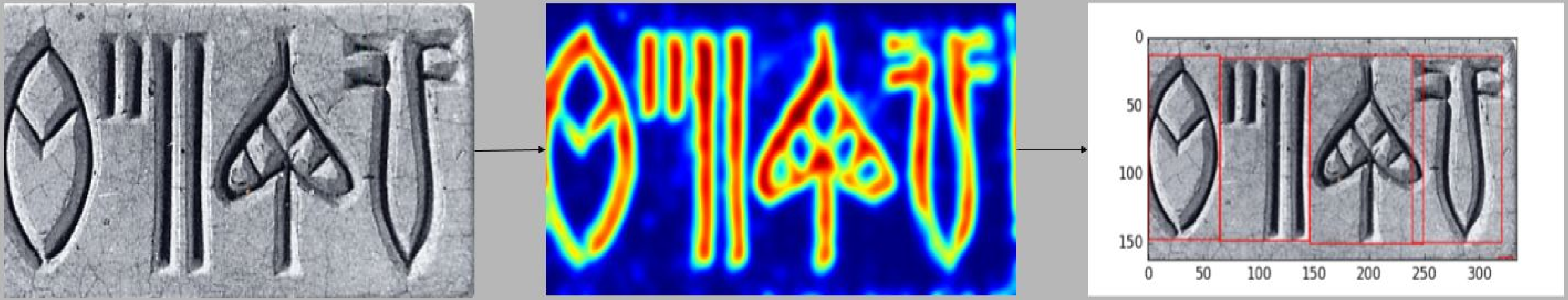
Text Region Formulation

Builds the text-only regions from the labeled region proposals (two level hierarchy)

- Draw TextBox
    - Merge Two “text” regions or a “text” region and a “both” region that are aligned along the same horizontal or vertical axis => TextBox
- Trim TextBox
    - Clip off the non-textual information (“no-text”) in those pairs of region proposals, where a “text box”/“text” region and a “no-text” region were overlapping

Symbol Segmentation

Segments out the individual graphemes from the precise text-only region proposals.

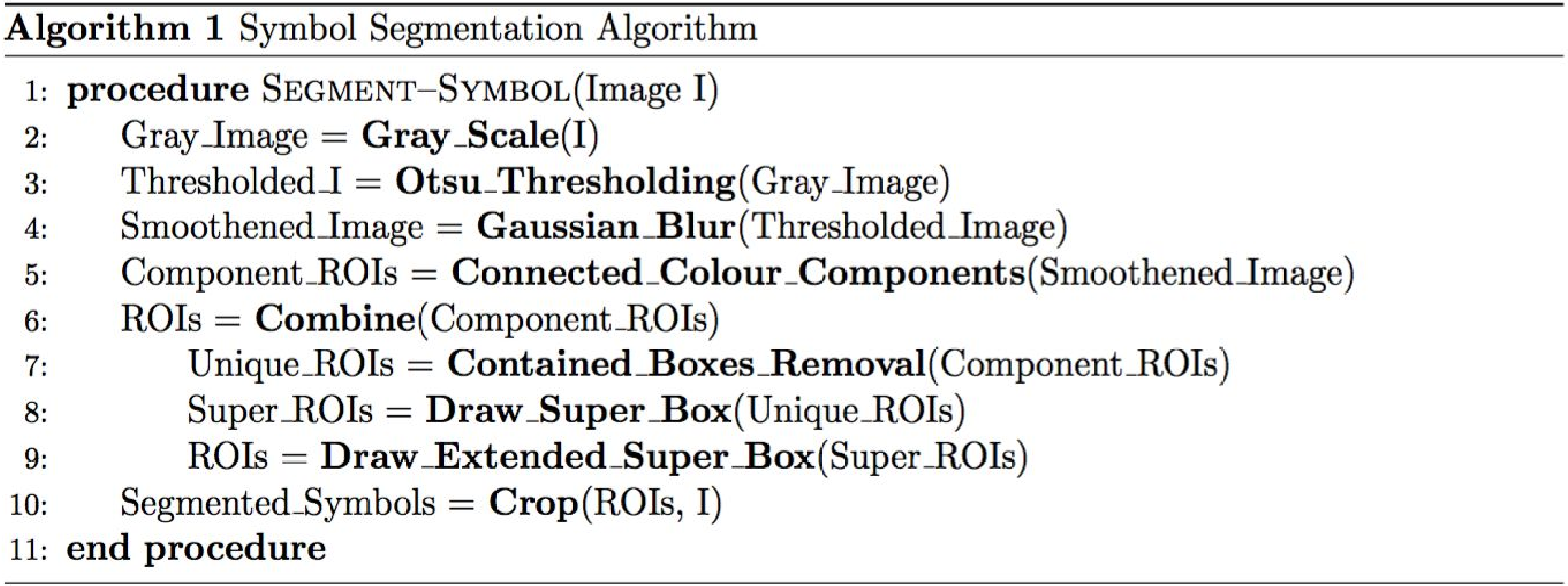


Text Region

1.Gray Scale
2.Otsu Thresholding
3.Gaussian Blur

4.Connected Colour Components
5.Combine Component Regions
6.Crop ROIs

Symbol Segmentation - The Algorithm

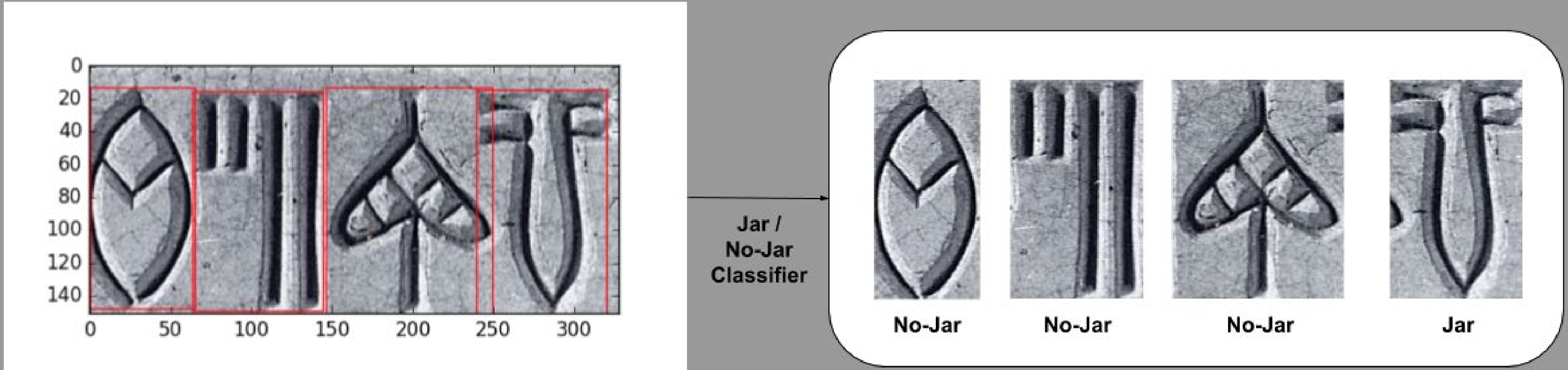


Algorithm 1 Symbol Segmentation Algorithm

```
procedure Segment–Symbol(Image I)
    Gray_Image = Gray_Scale(I)
    Thresholded_I = Otsu_Thresholding(Gray_Image)
    Smoothened_Image = Gaussian_Blur(Thresholded_Image)
    Component_ROIs = Connected_Colour_Components(Smoothened_Image)
    ROIs = Combine(Component_ROIs)
        Unique_ROIs = Contained_Boxes_Removal(Component_ROIs)
        Super_ROIs = Draw_Super_Box(Unique_ROIs)
        ROIs = Draw_Extended_Super_Box(Super_ROIs)
    Segmented_Symbols = Crop(ROIs, I)
end procedure
```

Symbol Identification

Takes individually cropped graphemes from the previous stage and classifies them into one of the 417 symbols (M77 Corpus)

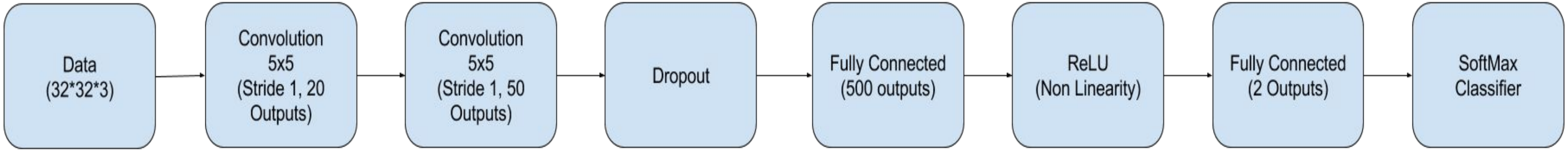


Symbol-wise Segmented Text Region

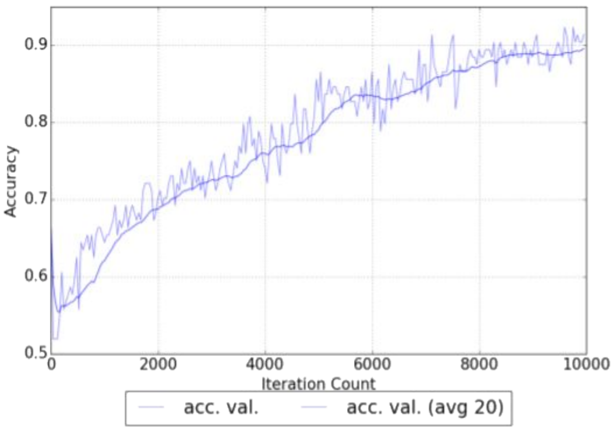
Classified Symbols

Symbol Identification - CNN Architecture

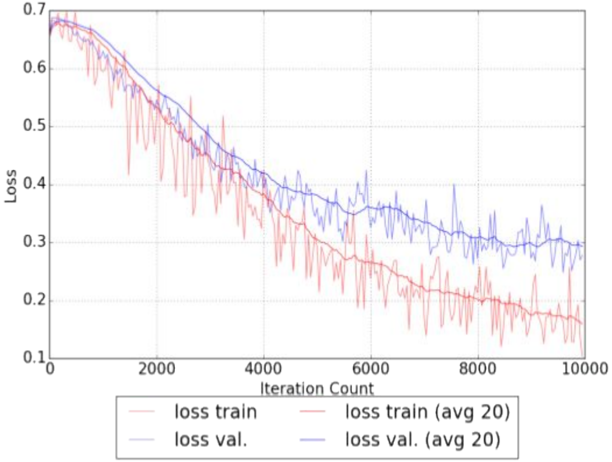
- Detects the presence or absence of the “Jar” sign - Binary classifier
- No transfer learning
- New architecture trained from scratch
- Accuracy score of **92.07%** when evaluated over the validation set of the “Jar-NoJar Dataset”


Data (32*32*3)
Convolution 5x5 (Stride 1, 20 Outputs)
Convolution 5x5 (Stride 1, 50 Outputs)
Dropout
Fully Connected (500 outputs)
ReLU (Non Linearity)
Fully Connected (2 Outputs)
SoftMax Classifier

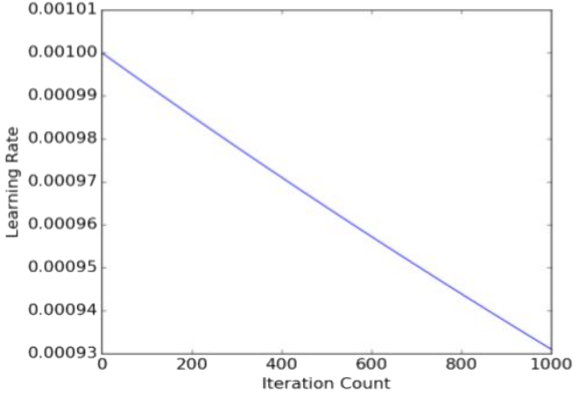
Symbol Identification CNN's Graphs



(a)

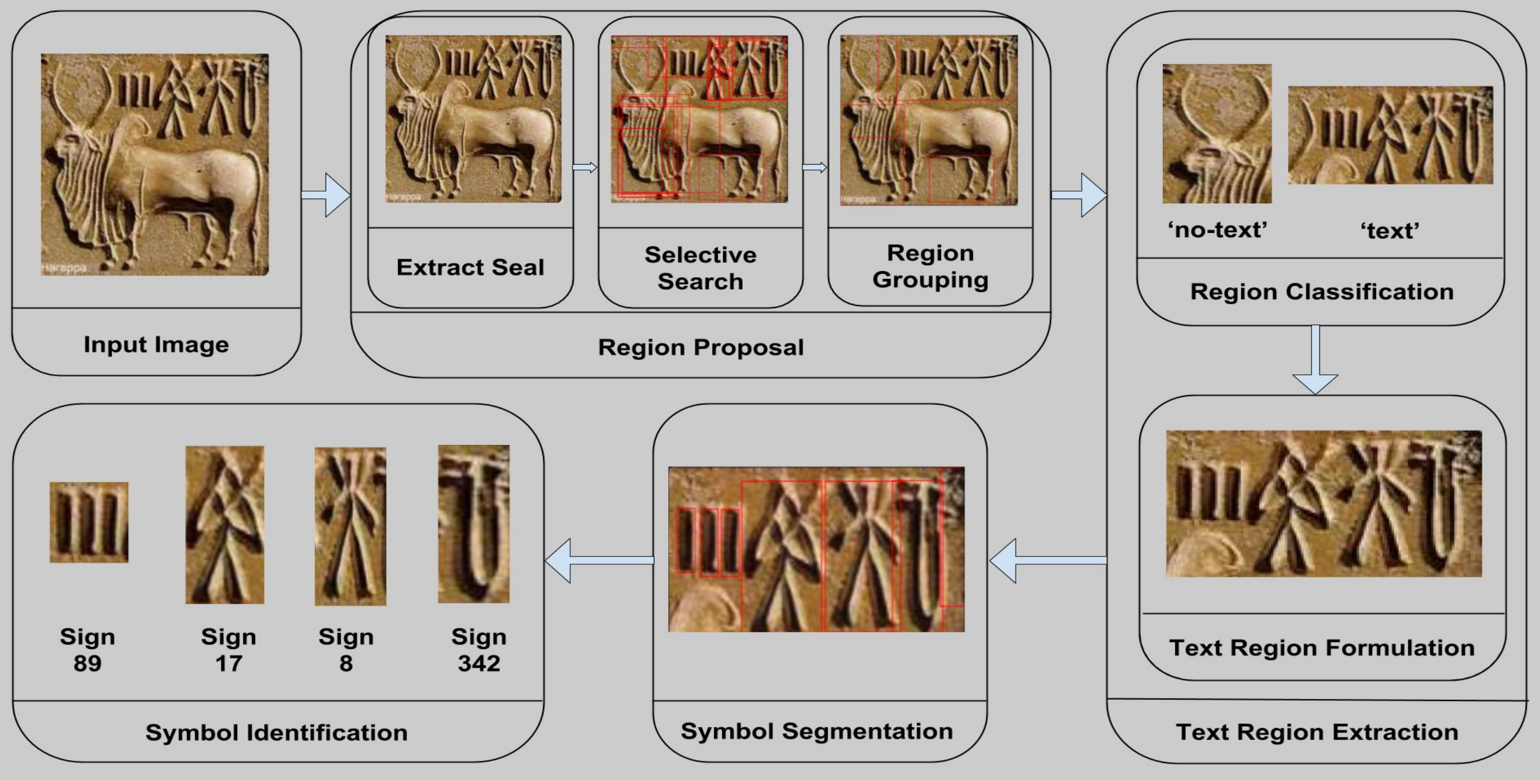


(b)

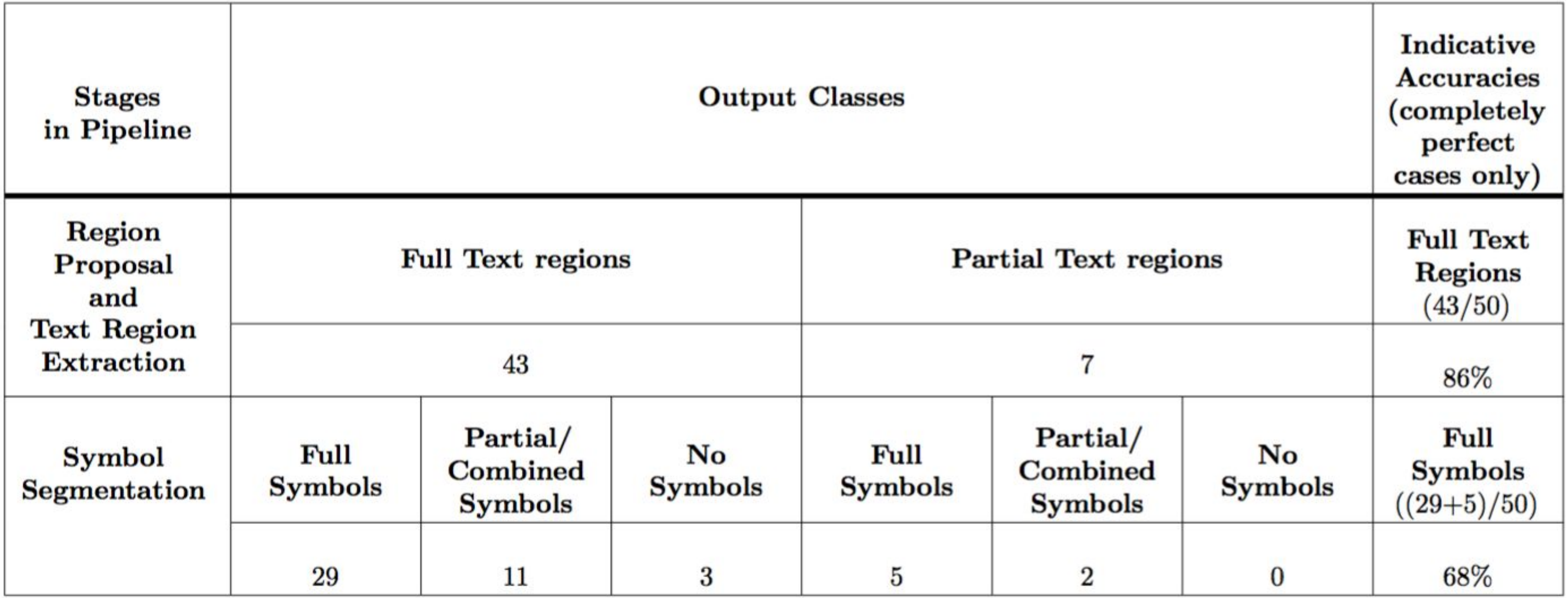


(c)

An Example Flow



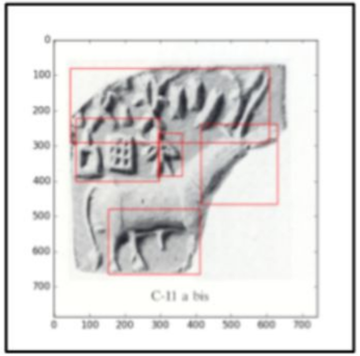
Evaluating the Pipeline

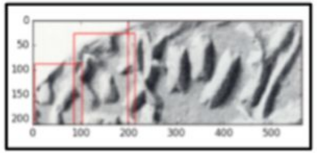


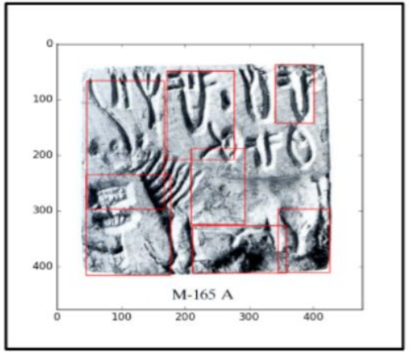
Stages in Pipeline	Output Classes						Indicative Accuracies (completely perfect cases only)
Region Proposal and Text Region Extraction	Full Text regions			Partial Text regions			Full Text Regions (43/50)
	43			7			86%
Symbol Segmentation	Full Symbols	Partial/ Combined Symbols	No Symbols	Full Symbols	Partial/ Combined Symbols	No Symbols	Full Symbols ((29+5)/50)
	29	11	3	5	2	0	68%

Limitations

- Non ML based techniques used in Region Proposal and Symbol Segmentation stages
  - Leads to some parts of the text region and symbols getting missed
- Extremely damaged seals can never be fully read
- Very complex grapheme arrangement structure
- Limited to identify only few frequent symbols of the 417
  - Issue is with the availability of such a dataset, it needs to compiled







Generalizing to Other Ancient Inscriptions

- We need
  - Sufficient amounts of data in the specified format for the new inscription
  - Tweak the region grouping and symbol segmentation stages and their parameters to harness the new inscription’s ergonomics

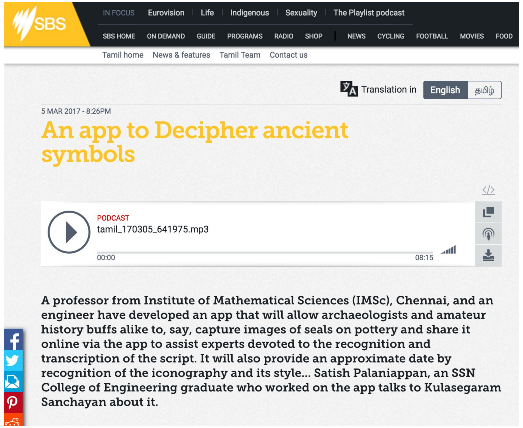
The Buzz !

CIPHER WAR

After a century of failing to crack an ancient script, linguists turn to machines

by Mallory Locklear | Jan 25, 2017, 9:00am EST

Meanwhile, Ronojoy Adhikari, a physics professor at The Institute of Mathematical Sciences in Chennai, India, and his research associate Satish Palaniappan are working on a program that can accurately extract symbols from a photo of an Indus artifact. "If an archaeologist goes to an Indus site and finds a new seal, it takes a lot of time for those seals to actually be mapped and added to a database if it's done manually," says Palaniappan. "In our case the ultimate aim is just with a photograph of a particular seal to be able to extract out the text regions automatically." He and Adhikari are working on building an app that archaeologists can bring to a site on a mobile device that will extract new inscriptions instantly.



5 MAR 2017 - 8:26PM

An app to Decipher ancient symbols

tamil_170305_641975.mp3

A professor from Institute of Mathematical Sciences (IMSc), Chennai, and an engineer have developed an app that will allow archaeologists and amateur history buffs alike to, say, capture images of seals on pottery and share it online via the app to assist experts devoted to the recognition and transcription of the script. It will also provide an approximate date by recognition of the iconography and its style... Satish Palaniappan, an SSN College of Engineering graduate who worked on the app talks to Kulasegaram Sanchayan about it.

Researchers Look To Widen Script Database, Solve Mystery

New app could help decipher ancient Indus Valley symbols

U.Tejonmayam
@timesgroup.com

As the Egyptian civilisation flourished and its calligraphers documented the rise and fall of one of man's greatest cultures in the period circa 3,500BC-1,300BC, another great civilisation arose in the Indus Valley in the northwest of the Indian subcontinent.

Much less is known about the Indus Valley Civilisation than its Egyptian equivalent, however — about its development, governance, activities, discoveries and daily life — because historians are comparatively short on information and yet to fully interpret the script of these ancient people.

To the common man, however, the limited corpus of hieroglyphs and other symbols that historians have uncovered from the Indus Valley bear an uncanny resemblance to those found along River Nile. Scientists linked the script to ancient Dravidian languages and an early form of Sanskrit but its meaning remains an enigma.

Here's a discovery, however, that could help change that.

Artificial intelligence involved in innovations like self-driving cars, which mimics the functions of the human brain, may now aid researchers to develop a computerised database for Indus script that could eventually help decipher the texts. Scientists are also working on a mobile application of the software.

The technology will allow archaeologists and amateur history buffs alike to, say, capture images of seals on pottery and share it online via the app to assist experts devoted to the recognition and transcription of the script. It will also provide an approximate date by recognition of the iconography and its style. The app will filter the text from the image and identify the presence or absence of individual characters in an existing database. If it is a known symbol, the app will display a number representing each character in the texts in the database; if not, it will include the symbol in the database. The output will be a string of graphemes (characters) and a corresponding number in the database, suitable for inclusion in a standard corpus.

A professor from Institute of Mathematical Sciences (IMSc), Chennai, and an engineer developed the app. Satish Palaniappan, an SSN College of Engineering graduate who worked on the app, said the sequence of numbers may help in the search for similar sequences in script, giving researchers a chance to draw inferences like origins or link between regions of the Indus Valley.

"That will be our next stage of research," Palaniappan said. "This app is for data accumulation and to learn new symbols. Automating corpus preparation will speed up research to decipher the script."

IMSc professor Ronojoy Adhikari said the app will augment the available corpus of Indus texts by automatically transcribing writing on artefacts. "There will be more texts to study," he said.

Researchers used 'deep learning' to develop the technology. "Deep learning is basically an artificial neural network-based learning technique, inspired by layers of interconnected neurons in the brain that interact and make decisions," Palaniappan said. "Deep learning has never been used before in epigraphic research."

The app is crucial to make big leaps in epigraphic research, "A researcher has to know the history and sequence of symbols," Palaniappan said. "It takes years to compile texts from artefacts and put them in a form that a computer will understand. We wanted to bridge that gap."

DAWN OF TIME

The computer application can be used to identify elements belonging to the Indus script

› **The image of the seal is scanned on all sides that are likely to have the Indus script symbols** or depictions of animals like bulls and unicorns, and deities

› The image is further classified into 'text', 'no-text', and 'both'. **The 'text' region consists of only the Indus script graphemes (unit or letter), the 'no-text' region consists of non-graphemic elements**

› A customised algorithm **segments out individual units or letters to identify it**

› **The identitified units or letters are classified into one of 417 classes, the known number of Indus graphemes,** according to scholar Iravatham Mahadevan

› **The results are in a sequence of numbers.** New characters are added to the database

ASI recently excavated structures resembling those at **Harappa from Keeladi village, Sivagangai district. They are believed to be 2,000 years old**

MODERN TAKE ON HISTORY

› **The Indus Valley Civilisation is one of the oldest known of the ancient civilisations.** It existed from 2500 BC to 1300 BC

› Discovered in the 1920s, it was spread over **12.6lakh sqkm covering parts of Pakistan, Afghanistan and India**

› It included around **1,000 well-planned villages, towns and cities, with two prominent cities, Harappa and Mohenjo-daro**

› Seals were used for **ritualistic, commercial and religious purposes**

› **Copper, bronze, pottery and terracotta toys, steatite (soapstone) seals** were retrieved from the sites. Many of these were engraved with animal figures

Though the origin and decline of the **civilisation remain a mystery, there are theories that floods, deforestation or the invasion of Aryans** could be among reasons

Chennai team taps AI to read Indus Script

The algorithm uses 'deep neural networks' which are also used in self-driving cars

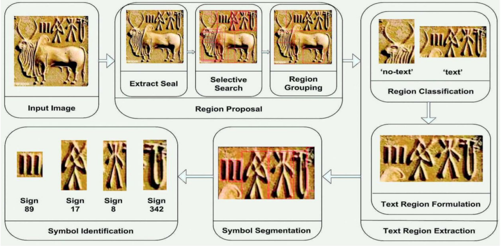
SHUBASHREE DESIKAN

The Indus script has long challenged epigraphists because of the difficulty in reading and classifying text and symbols on the artefacts. Now, a Chennai-based team of scientists has built a programme which eases the process.

Ronojoy Adhikari of The Institute of Mathematical Sciences and Satish Palaniappan, who is at Sri Sivasubramaniya Nadar College of Engineering, have developed a "deep-learning" algorithm that can read the Indus script from images of artefacts such as a seal or pottery that contain Indus writing.

Scanning the image, the algorithm smartly "recognises" the region of the image that contains the script, breaks it up into individual graphemes (the term in linguistics for the smallest unit of the script) and finally identifies these using data from a standard corpus. In linguistics the term corpus is used to describe a large collection of texts which, among other things, are used to carry out statistical analyses of languages.

The algorithms come under a class of artificial intelligence called "deep neural networks." "These have been a major part of the game-changing technology behind self-driving cars and Go-playing bots that surpass human performance," says Satish Palaniappan. The deep neural network mimics the working of the mammalian visual cortex, known as convolutional neural network (CNN), which breaks the field into overlapping regions. The features found in each region are hierarchically combined by the network to build a composite understanding of the whole picture.



Step by step: Scanning the image, the algorithm smartly processes the data in three steps to place its elements within the standard corpus. SPECIAL ARRANGEMENT

The process consists of three phases: In the first phase, the input images are broken into sub-images that contain graphemes only, by trimming out the areas that do not have graphemes. The grapheme-containing areas are further trimmed into single-grapheme pieces. Lastly, each of these single graphemes is classified to match one of the 417 symbols discovered so far in the Indus script.

Indus script

The Indus valley script is much older than the Prakrit and Tamil-Brahmi scripts. However, unlike the latter two, it has not yet been deciphered because a bilingual text has not yet been found.

A bilingual text has in many other cases aided archaeologists in understanding ancient scripts, for example, the Rosetta stone. This stone which was found in the eighteenth century carries inscriptions of a decree, issued in 196 BCE, in three parts, the first two in ancient Egyptian hieroglyphic and the Demotic scripts, while the bottom is in Ancient Greek. Since the decree was the same, the Rosetta stone provided the key to deciphering Hieroglyphs. For the lack of such a "Rosetta stone," the Indus script remains undeciphered today.

It is a major effort to even build a standard corpus of the language and decode the writing on existing artefacts and map them to this standard corpus. The most widely accepted corpora of Indus scripts was brought together by the efforts of Iravatham Mahadevan, noted Indian epigraphist, from the 3,700 texts and 417 unique signs collected so far.

When asked about the relevance of this work, Dr Mahadevan says, "It [the algorithm] represents a significant advance in the computerised study of the Indus Script. I wish I had this software 40 years ago when I compiled the Indus concordance."

Dr Mahadevan says, "It **[the algorithm]** represents a significant advance in the computerised study of the Indus Script. I wish I had this software 40 years ago when I compiled the Indus concordance."

Thank You!

Satish Palaniappan