Extracting Textual Information from Images

Dr. Kunal Banerjee

Principal Data Scientist

8-May-2024

Walmart

Brief Biography

- Professional Experience
  - **Walmart:** Principal Data Scientist (Sep 2020 – Present)
  - **Intel Labs:** Research Scientist (Aug 2015 – Aug 2020)
- Educational Background
  - **PhD:** Computer Science & Engineering, IIT Kharagpur (2010 – 2015)
  - **BTech:** Computer Science & Engineering, Heritage Inst of Tech (2004 – 2008)
- Research Highlights
  - 8 journal publications
  - More than 50 conference/workshop publicatons
  - 1078 citations (as on May 9, 2024)
  - Best PhD Thesis Awards, Best Paper Awards, Special Mentions
  - IEEE Senior Member, ACM Senior Member

Publications Covered in this Talk

- ★ P Dugar et al., “From Pixels To Words: A Scalable Journey Of Text Information From Product Images To Retail Catalog,” CIKM 2021
- ★ P Dugar et al., “Don’t Miss the Fine Print! An Enhanced Framework To Extract Text From Low Resolution Images,” VISAPP 2022
- ★ S Misra et al., “Designing a Vision Transformer based Enhanced Text Extractor from Product Images,” CoDS-CoMAD 2023
- ★ S Misra et al., “BARGAIN: A Super-Resolution Technique to Gain High-Resolution Images for Barcodes,” CoDS-CoMAD 2024

CIKM 2021

Applied Research Paper Track CIKM '21, November 1–5, 2021, Virtual Event, Australia

From Pixels To Words: A Scalable Journey Of Text Information From Product Images To Retail Catalog

Pranay Dugar*
Walmart Global Tech
Bangalore, Karnataka, India
pranay.dugar@walmart.com

Rajesh Shreedhar Bhat*
Walmart Global Tech
Bangalore, Karnataka, India
rajesh.bhat@walmart.com

Asit Sharad Tarsode
Walmart Global Tech
Bangalore, Karnataka, India
asit.sharad.tarsode@walmart.com

Uddipto Dutta
Walmart Global Tech
Bangalore, Karnataka, India
uddipto.dutta@walmart.com

Kunal Banerjee
Walmart Global Tech
Bangalore, Karnataka, India
kunal.banerjee1@walmart.com

Anirban Chatterjee
Walmart Global Tech
Bangalore, Karnataka, India
anirban.chatterjee@walmart.com

Vijay Srinivas Agneeswaran
Walmart Global Tech
Bangalore, Karnataka, India
vijay.agneeswaran@walmart.com

Text Extraction from Images: Challenges

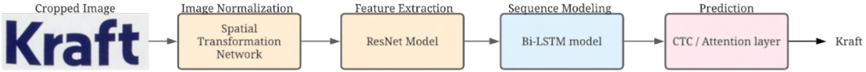
Characteristics of texts on product images:

- Non-standard fonts and sizes
- Text can be vertical or inverted
- Text can be irregularly oriented or curved
- Non-dictionary words (e.g., brand names)
- High local entropy

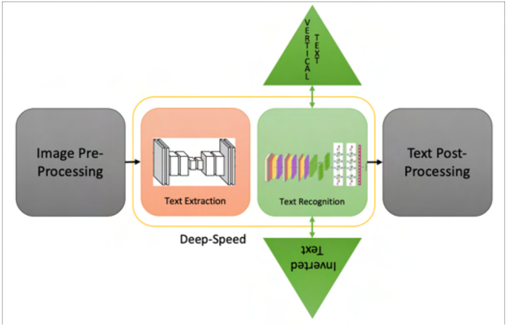
Text Extraction from Images: Use-cases

- In the context of scale at which Walmart operates, the text from an image can be **a richer and more accurate source of data** than human inputs
- Used in *several applications* such as Attribute Extraction (MRP, Country of Origin, Ingredients), Offensive Text Classification, Product Matching
- The solution provided is proven to work at million image scale for various retail business units within Walmart while saving 30% computational cost in both the training and the inference stages

Generic Pipeline for Text Extraction from Images



Walmart's Pipeline for Text Extraction from Images



Handling Vertical Texts

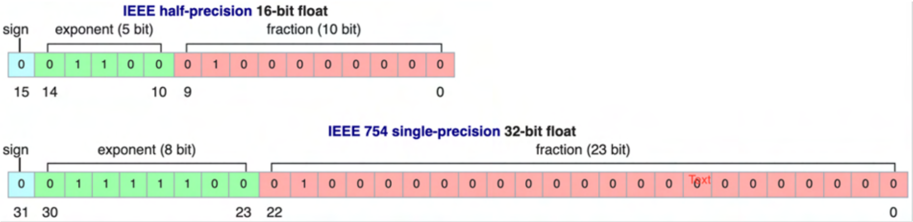
- Case 1 can be handled by rotating the text by 90° or 270°
- Case 2 requires slicing each character and then putting these together to get the word

Handling Inverted Texts

Confidence scores are **low**

Confidence scores are **high**

High-Performance Computing with DeepSpeed

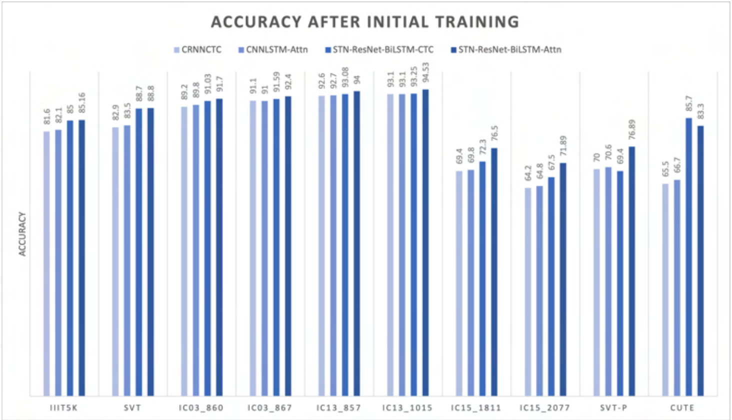


- Microsoft's DeepSpeed is a library for scaling DL training and inference
- DeepSpeed has been utilized on two fronts:
  - Data Parallelism
  - Low Precision Training
- Allows for a reduced training time for the model
- Reduces the latency of model at inference time with negligible drop in accuracy
- Since the model fits within memory of a single GPU, only data parallelism was explored, not model and pipeline parallelisms

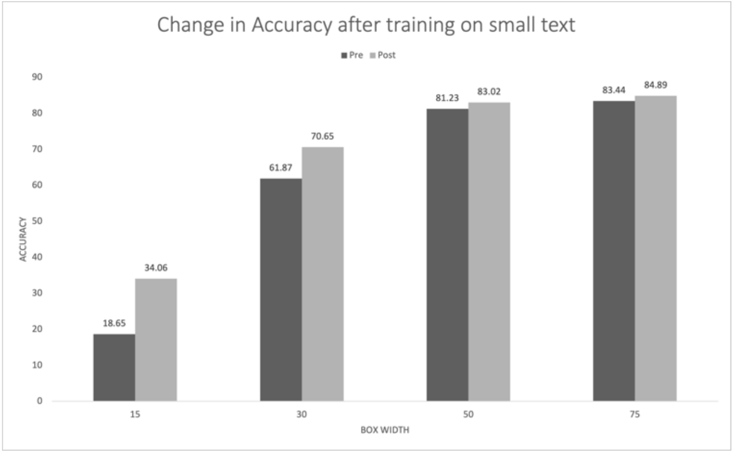
Datasets

- Our main source of training data was the Walmart catalog text imprinted onto a background using SynthText
- Furthermore, we used various open-source datasets: IIIT5K, SVT, SVT-P, ICDAR03, ICDAR13, ICDAR15 and CUTE – consisting of both normal and arbitrary shaped text
- We further incorporated a data augmentation approach where the text images were resized to much smaller sizes to improve the accuracy on smaller text

Initial Results



Improvements on Small Text



Text Extraction from Images in Action: Cargo Identification

VISAPP 2022

Don't Miss the Fine Print! An Enhanced Framework to Extract Text from Low Resolution Images

Pranay Dugar[1], Aditya Vikram[2,*], Anirban Chatterjee[1], Kunal Banerjee[1][a] and Vijay Agneeswaran[1]
[1]*Walmart Global Tech, Bangalore, India*
[2]*Flipkart, Bangalore, India*

Low-Resolution Images can be Challenging

- Text extraction models perform impressively on clear texts but show a significant decline in accuracy when recognizing text in low-resolution images
- Off-the-shelf super-resolution tools produce images that *appear* sharper but the texts often stay illegible

Contributions

- An approach to generate synthetic LR-HR paired data that is generalizable to real case scenarios for product images
- A variation of perceptual loss termed recognition loss
- An improvised multi-loss function composed of detection and recognition losses as well as image features
- Visually and analytically superior results

Data Generation and Annotation

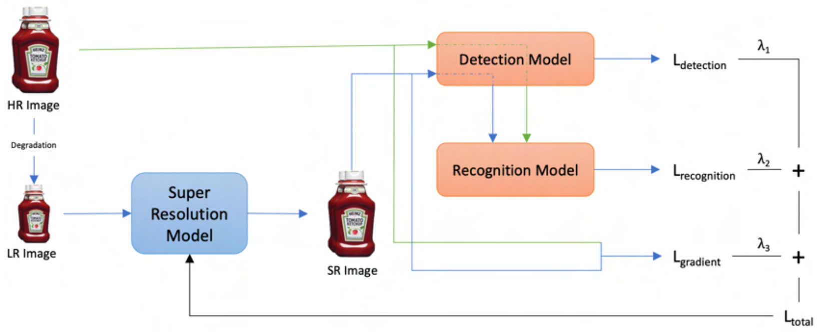
Our proposal involves a two-stage method that generates a synthetic dataset from natural scene text datasets that is more robust to image distortions

1. Downsample by 4x using a randomly-chosen interpolation technique: linear, bicubic, nearest-neighbor, etc.
2. Upsample by 2x using a randomly-chosen extrapolation technique: linear, bicubic, nearest-neighbor, etc.

☞ The different randomly-chosen techniques for both downsampling and upsampling introduce more randomness

☞ The randomness translates into robustness during subsequent training

Text Super-Resolution Model Architecture



Loss Functions

- **Gradient Loss:** to allow model to better detect edges, gradient is calculated along each channel, followed by the mean across channels to negate abnormalities across different image channels

$$L_{grad} = ||\Delta I^{HR} - \Delta I^{SR}||_1$$

$$\Delta I = \frac{1}{2 \times channels} \sum_{channels} (\delta I_{width} + \delta I_{height})$$

- **Recognition Loss:** uses feature maps generated by the fourth convolutional block of the pre-trained encoder of the text recognition ASTER model to compute the perceptual loss

$$L_{rec} = ||\Psi_n(I^{HR}) - \Psi_n(I^{SR})||_2$$

- **Detection Loss:** to ensure that the model can detect the precise locations of all the texts in an image with higher accuracy, we use the predicted coordinates of the SR and the HR images to create two mask images consisting of detected regions being masked out and compute the loss using these masks

$$img\_mask(p) = \begin{cases} 1, & \text{if p in detected box} \\ 0, & \text{otherwise} \end{cases}$$

$$L_{det} = \frac{1}{P} \sum ||HR\_mask(p) - SR\_mask(p)||^2$$

Performance Metrics

- **PSNR:** ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation; in case of images, each pixel can be considered as a component of a signal with 8-bit RGB values
- **SSIM:** it's a measure that tries to replicate the way human visual system (HSV color model) works; it is designed based on three factors: correlation, luminance distortion and contrast distortion
- **Accuracy:** exact match between the ground truth word and the predicted word
  E.g. GT: "Salt", Prediction: "Salt" $\rightarrow$ Accuracy: 1.0
  GT: "Salt", Prediction: "Sale" $\rightarrow$ Accuracy: 0.0
- **Normalized Edit Distance:** A fuzzy match between the ground truth word and the predicted word
  E.g. GT: "Salt", Prediction: "Salt" $\rightarrow$ NormED: 1.0
  GT: "Salt", Prediction: "Sale" $\rightarrow$ NormED: 0.75

Datasets

- We performed experiments on datasets designed for the task of text extraction from images
- Open-source datasets: ICDAR2013, ICDAR2015 and SVT
- They provide word-level ground truth boxes of text
- We use these ground truth boxes as the area of consideration for analytical scoring metrics
- Two ways to gauge the performance of a model: visual perception and analytical scores

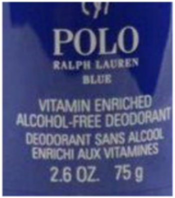
Visual Perception Results

ESRGAN

IMDN

DNCNN

Our Model



HR Image

Results based on Performance Metrics

Model	ICDAR2013		ICDAR2015		SVT	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
ESRGAN	29.432	0.827	29.338	0.826	30.458	0.839
IMDN	**32.266**	0.881	**32.170**	0.881	**33.383**	0.895
DNCNN	32.022	**0.897**	32.017	**0.897**	32.464	0.910
Our Model	29.236	0.882	29.122	0.881	32.545	**0.928**

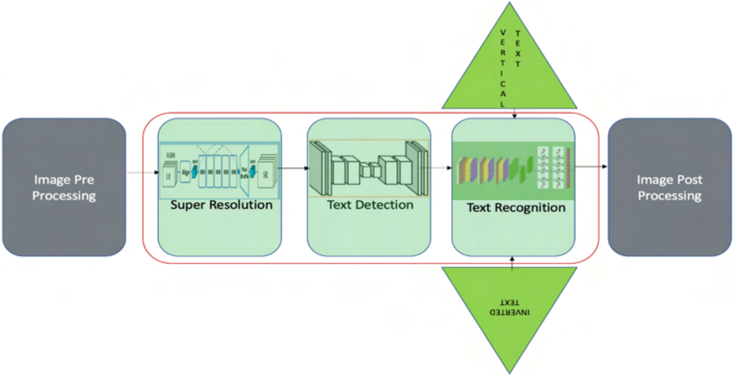
Model	ICDAR2013		ICDAR2015		SVT	
	Accuracy	NormED	Accuracy	NormED	Accuracy	NormED
ESRGAN	0.808	0.881	0.814	0.905	0.684	0.817
IMDN	0.833	0.919	0.836	0.938	0.721	0.848
DNCNN	0.853	0.919	0.863	0.945	0.726	0.853
Our Model	**0.876**	**0.928**	**0.890**	**0.958**	**0.821**	**0.921**

CoDS-CoMAD 2023

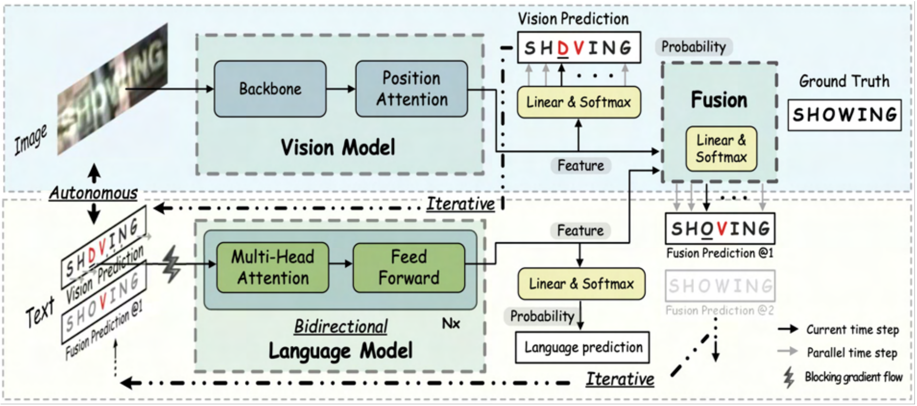
Designing a Vision Transformer based Enhanced Text Extractor for Product Images

Saptarshi Misra, Pranay Dugar, Anirban Chatterjee, Lalitdutt Parsai, Kunal Banerjee
{saptarshi.misra,pranay.dugar,anirban.chatterjee,lalitdutt.parsai,kunal.banerjee1}@walmart.com
Walmart Global Tech
Bangalore, India

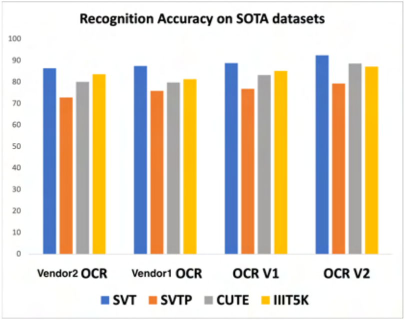
Text Extraction Pipeline: Recap

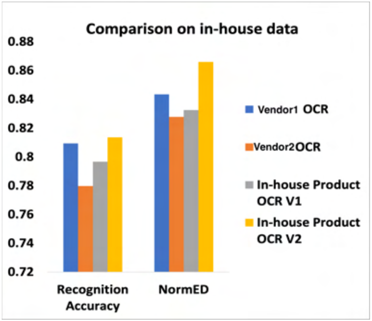


ABINet Model (Source: Fang et al., CVPR 21)



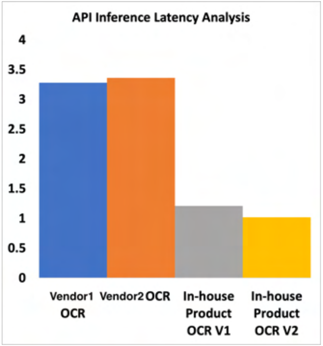
Results on public and in-house datasets





Inference Latency Comparison

- Experiments carried out on Intel® Xeon® CPU @ 2.30GHz connected to Nvidia Tesla V100-SXM2 GPUs
- Lesser latency is better



CoDS-CoMAD 2024

BARGAIN: A Super-Resolution Technique to Gain High-Resolution Images for Barcodes

Saptarshi Misra
saptarshi.misra@walmart.com
Walmart Global Tech
Bangalore, India

Kunal Banerjee
kunal.banerjee1@walmart.com
Walmart Global Tech
Bangalore, India

Anirban Chatterjee
anirban.chatterjee@walmart.com
Walmart Global Tech
Bangalore, India

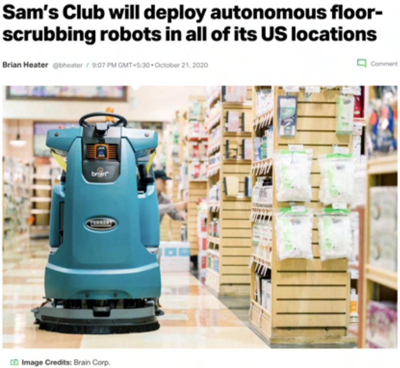
Inventory Management using Images

Sam's Club will deploy autonomous floor-scrubbing robots in all of its US locations

Brian Heater @bheater / 9:07 PM GMT+5:30 • October 21, 2020 Comment

Image Credits: Brain Corp.

Inventory Management using Images



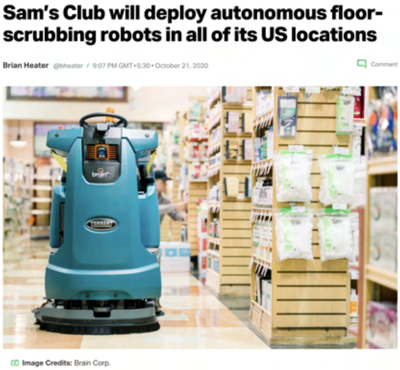
Sam's Club will deploy autonomous floor-scrubbing robots in all of its US locations

Brian Heater @bheater / 9:07 PM GMT+5:30 • October 21, 2020

Comment

Image Credits: Brain Corp.

Inventory Management using Images

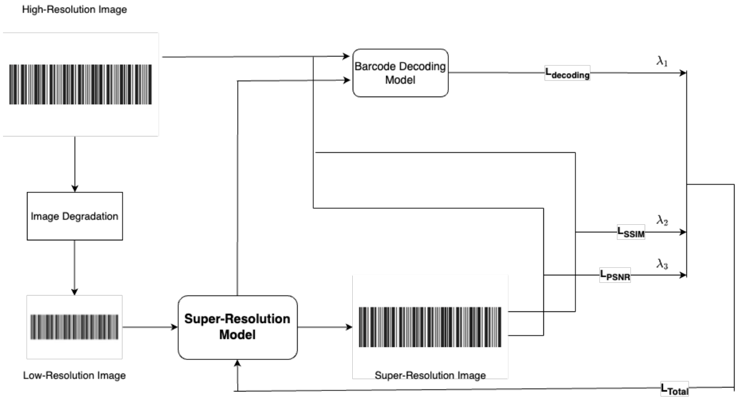

Sam's Club will deploy autonomous floor-scrubbing robots in all of its US locations

Brian Heater @bheater / 9:07 PM GMT+5:30 • October 21, 2020 Comment

Image Credits: Brain Corp.

- Text extraction can't distinguish between similar SKUs (Pepsi 1l vs Pepsi 1.2l)
- Barcodes are unique for each SKU

Super-Resolution for Barcodes



Experimental Results

Improvement in barcode decoding accuracy and ablation study

Image Type	Number of images decoded by ZXing
Low-resolution images	14000
High-resolution images generated using bicubic interpolation	14047
High-resolution images generated using bilinear interpolation	14029
High-resolution images generated using nearest neighbour interpolation	14035
High-resolution images generated using ESRGAN [5]	14125
High-resolution images generated using IMDN [6]	14097
High-resolution images generated using DNCNN [7]	14163
High-resolution images generated using the SR model proposed in [2]	14200
High-resolution images generated using ESRT transformer model [3]	14930
High-resolution images generated using EMT transformer model [4]	14976
High-resolution images generated using our model BARGAIN	**15800**
Ablation studies on the proposed Loss Function	
High-resolution images generated using our model BARGAIN and SSIM + PSNR loss	14947
High-resolution images generated using our model BARGAIN and PSNR + Barcode Decoding loss	15437
High-resolution images generated using our model BARGAIN and SSIM + Barcode Decoding loss	15623

Impact of different resolutions and at different camera distances

Barcode image resolution in MP	No. of barcodes detected at =1 m	No. of barcodes detected at =1 m on SR	% improv	No. of barcodes detected at =1.1m	No. of barcodes detected at =1.1m on SR	% improv	No. of barcodes detected at =1.2m	No. of barcodes detected at =1.2m on SR	% improv	No. of barcodes detected at =1.3m	No. of barcodes detected at =1.3m on SR	% improv
4	4203	6125	45.73	4153	6097	46.81	2057	5453	165	1096	3119	184.84
6	4105	7721	88.09	4155	6067	46.02	2272	6384	180.96	1532	3095	102.02
8	6359	6418	0.93	5091	5172	1.59	4009	5741	43.20	2067	4059	96.37
10	6172	7251	17.48	6831	6952	1.77	5164	5671	9.82	3098	4089	31.99
12	7091	7193	1.44	5176	6374	23.14	4432	6389	44.16	2987	5432	81.85

Conclusion

- Text embedded in images can be a rich source of information
- Text extraction can help in catalog enrichment, product matching, profanity detection, etc.
- Text orientation, curvature of surfaces, bad lighting among others can pose challenges for text extraction
- Text focused super-resolution as a pre-processing step can help (AFAIK no earlier SR technique specialized on a specific feature in an image)
- Identifying barcodes can further help in inventory management
- Our prescribed models and methodologies improve upon the SOTA and contribute towards multiple Walmart's businesses
- Further improvements can be achieved by adding use-case specific dictionaries (in different languages) as a post-processing step
- We are still far from decoding the highly degraded low-resolution scene texts, and the field requires more effort to solve the same

Thank you!

Email: Kunal.Banerjee1@walmart.com
Homepage: https://kunalbanerjee.github.io/