Advanced Machine Learning Project

Final Presentation

Text detection from images using deep learning

Team NPM :
Neeraj Kumar, Parag Jain, Mohsen

Problem Statement

- Quebec - predominantly French speaking and writing province
- Annual international student arrival for education
- Difficulty navigating without any external help

Motivation

- Personal experience

- Purchasing groceries, ordering food in restaurants - a major challenge

Solution to the Problem

- Service to translate text from one language to another.

- Using image with text as input, 3 step solution :

    1. Text extraction from input image
    2. Text recognition from extracted text
    3. Conversion of text from one language to another

Scope of the project

Text extraction from image

Text extraction from images

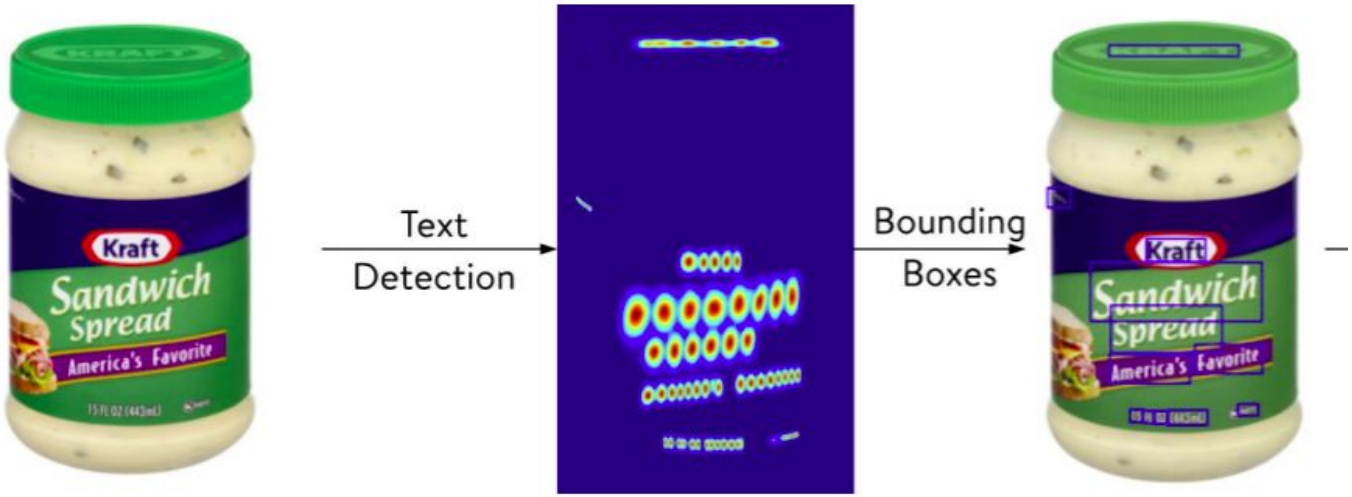
- Using DIP - manual feature engineering
- Using deep learning - word-level annotated training set

Limitations

Arbitrary shaped Text

Curved Text

Solution diagram



Text detection using character-level annotations to

overcome the limitations of other approaches

What do we need?

- Model with the ability to :
  - localize individual character regions
  - link the detected characters to a text instance

- A way to quantitatively measure model's ability :
  - region score - indicating model's localization ability
  - affinity score - indicating model's linking ability

- Loss function - improve model's ability by means of providing feedback based on predicted scores

Let's take a look at the dataset

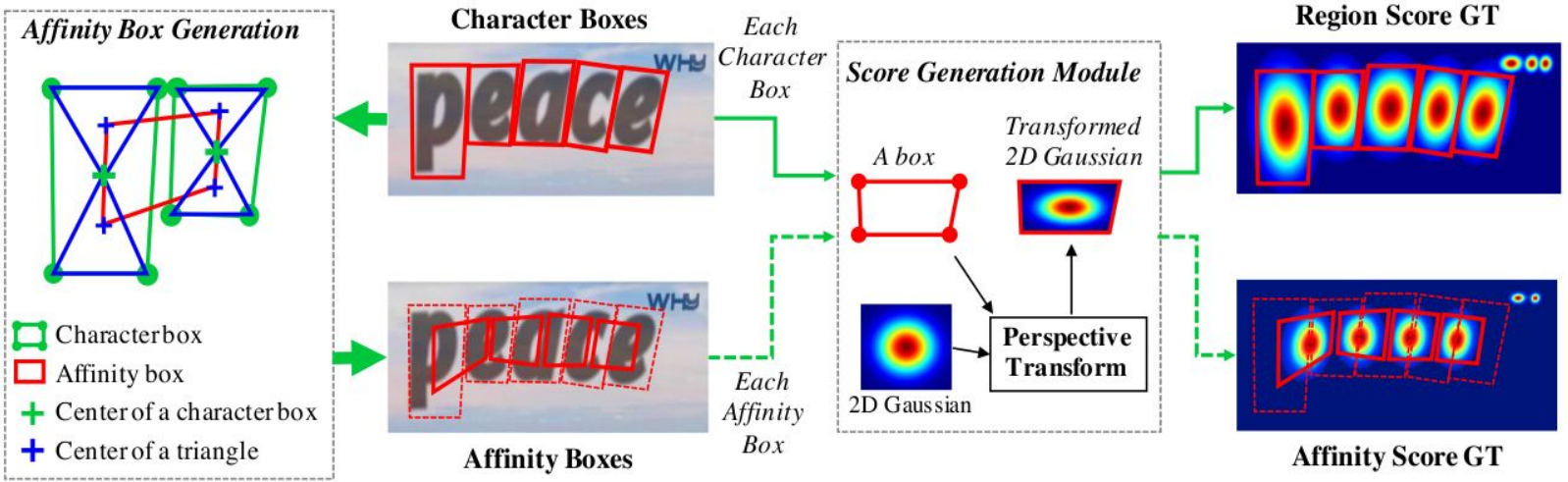
Training Dataset Overview

- Synthetic dataset :
  - character-level annotated text images
  - ~140000 images

Word	Character Level Bounding Boxes
Sender:	[121, 255], [145, 255], [121, 303], [145, 303]
Ecole	[140, 308], [173, 308], [140, 327], [173, 327]
Other words	[Top-left], [Top-right], [Bottom-right], [Bottom-left]

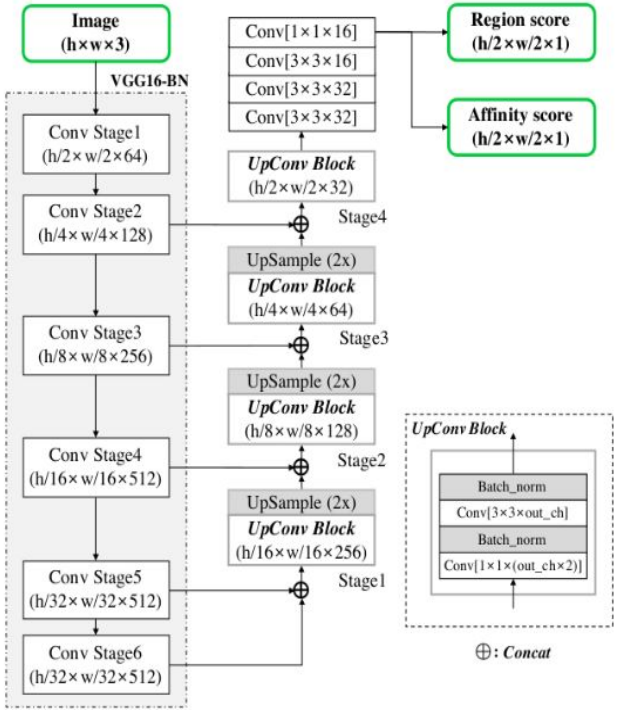
- Region-score and affinity score - to be derived

Deriving true labels using SynthText dataset



Model Architecture

VGG16 Network with Batch Normalization (Pre-trained)
+
U-net



Training loss for SynthText

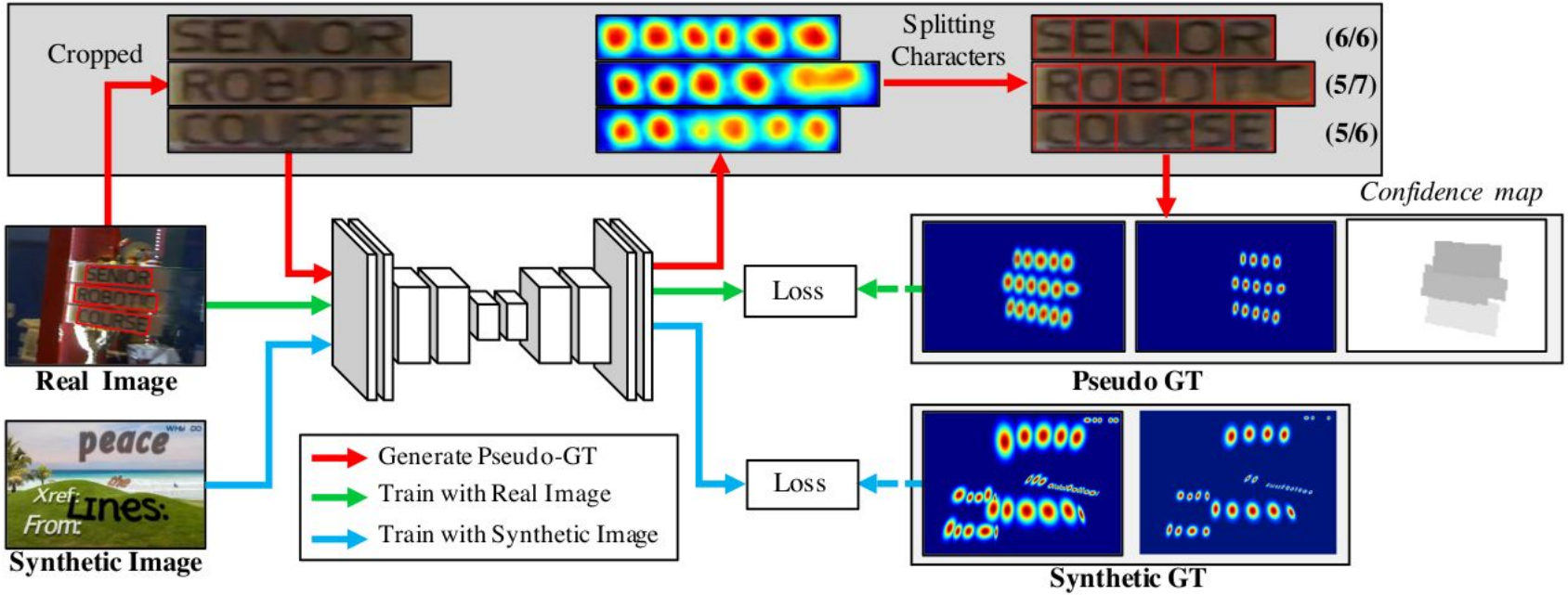
$$L = \sum_{p} S_c(p) \cdot \left(||S_r(p) - S_r^*(p)||_2^2 + ||S_a(p) - S_a^*(p)||_2^2 \right)$$

where, $S_r(p)$: Pixel value at any pixel p of the predicted region score
$S_r^*(p)$: Pixel value at any pixel p of the region score GT

$S_a(p)$: Pixel value at any pixel p of the predicted affinity score
$S_a^*(p)$: Pixel value at any pixel p of the affinity score GT

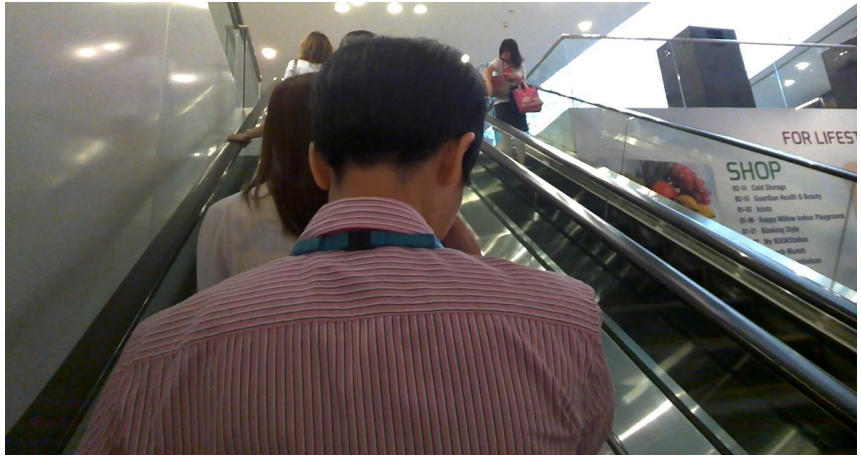
$S_c(p) = 1$

Training Procedure



Fine-tuning Dataset

- **ICDAR-2015/RR-2015**
  - Used to weakly supervise the network
  - ~1000 images



More Details on Loss Function

$$S_c(p) = \begin{cases} s_{\text{conf}}(w) & p \in R(w) \\ 1 & \text{otherwise} \end{cases}$$

- $R(w)$: bounding box region
- $l(w)$:the word length of the sample w
- $l^c(w)$: length of characters

$$s_{conf}(w) = \frac{l(w) - \min\left(l(w), |l(w) - l^c(w)|\right)}{l(w)}$$

where p denotes the pixel in the region $R(w)$. The objective L is defined as,

$$L = \sum_p S_c(p) \cdot \left(\|S_r(p) - S_r^*(p)\|_2^2 + \|S_a(p) - S_a^*(p)\|_2^2 \right)$$

where $S_r^*(p)$ and $S_a^*(p)$ denote the pseudo-ground truth region score and affinity map, respectively, and $S_r(p)$ and $S_a(p)$ denote the predicted region score and affinity score, respectively. When training with synthetic data, we can obtain the real ground truth, so $S_c(p)$ is set to 1 .

Performance Metrics

Recall	Precision	H-mean
84.3	89.8	86.9

CRAFT Paper

Recall	Precision	H-mean
79.39	85.5	81.14

Our Reproduction

Test Results

Inference Results

Ex 1 : Picture shot from mobile camera outside MILA campus

Ex 2 : Cookies box with description in French

Conclusion :

Regularization and the other Techniques

- **Adding a term $\|w - w_k\|^2$ to the loss function can be beneficial called proximal term.**
  - Encourages the learned parameter w to change gradually from its previous value w_k.
  - Prevent large fluctuations in the learned parameter values and potentially improve convergence.
  - Adding only the term $\|w\|^2$ to the loss function can lead to large changes in the learned parameter.
  - Slow down convergence and lead to poor generalization performance.
- Weight Initialization Technique:
  - Swish initialization [Based on the Swish activation function f(x) = x * sigmoid(x)]
  - Generalized Xavier [scales the variance of the Gaussian distribution]
- Technique Apply to the Activations:
  - Adaptive instance normalization [AdaIN transfer the style information of one image onto the content of another image, while preserving the content information.]
  - AdaIN improve the generalization and robustness of the CNN.
  - Cross-Layer Equalization[CLE: ensuring that the activations of each layer have the same distribution]

To improve in future

- Investigate ways to reduce the computational complexity, such as using **smaller network architectures** or **optimizing the inference process**.

- Explore ways to improve the robustness of the method to variations in lighting, image quality, and other factors that can affect **text detection performance.**

- **Evaluate on larger and more diverse datasets** to better understand its strengths and weaknesses and identify opportunities for further improvement.

Future work

- Complete the other 2 steps of the 3 part solution :
  - **Text recognition** from extracted text
  - **Text translation** from one language to another

- Make this solution available for developers to use as an API

Future Application : An Alternative Software To mathpix

The best image recognition for math and science

Mathpix is the only image to LaTeX converter with high-accuracy OCR features developed specifically for scientific documents like research papers

Math

$$\mathbf{P}[\mathcal{E}_2] = 1 - \prod_{j \in A} \mathbf{P}[Y_j = 0] = 1 - \prod_{j \in A} (1 - x_j/2).$$

$$\Gamma(z) = \int_0^\infty x^{z-1} \cdot e^{-x} dx$$

Text

Given the effective depth of our lightcones, the spatial resolution of the ELEPHANT suite, and taking into account that we consider only resolved halos, we can estimate that our catalogs will be spatially resolved down to $\sim 0.5 - 1h^{-1}$ Mpc [53]. Within the redshift range we use, this sets the minimum angular scales that we can consider as resolved to be $\theta_{res} \approx 0.05°$.

The only source of knowledge is experience.

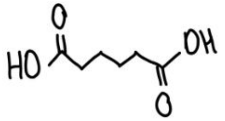
Tables

AAS	DM	LM	MDM	MLM	Discarded
237	146	130	23	188	275
238	35	34	6	52	127
	181	164	29	240	402

x	3	-5	8	4	0
y	3	-21	18	6	-6

Chemistry

$C_6H_{12}O_6 + O_2 \rightarrow CO_2 + H_2O$



Team contribution :

- Overall, project completion was achieved through group effort.

- Highlighting some of the initiatives taken by every individual in the team :

  - **Parag Jain**
    - Identifying the problem and preparing the proposal
    - Shortlisting of paper and literature review
    - Implementing the network architecture
  - **Neeraj Kumar**
    - GT label generation, pre-processing and implementing the training phase
    - Setting up the cluster and training
    - Keeping the team progress on track
    -
  - **Mohsen Dehghani**
    - Understanding and explaining the finer details of CRAFT
    - Testing and inference
    - Suggesting ways to improve model performance

Thank you!

Link to Project Code

Work related to the project can be found here.