DATA+AI
SUMMIT EUROPE
FORMERLY SPARK+AI SUMMIT

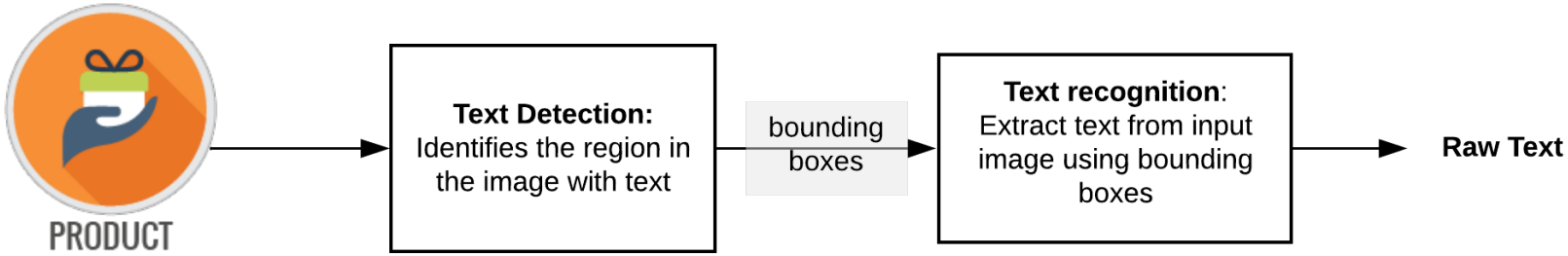
Detecting & Recognizing arbitrary shaped texts from Product Images

Rajesh Shreedhar Bhat
Senior Data Scientist, Walmart Global Tech India

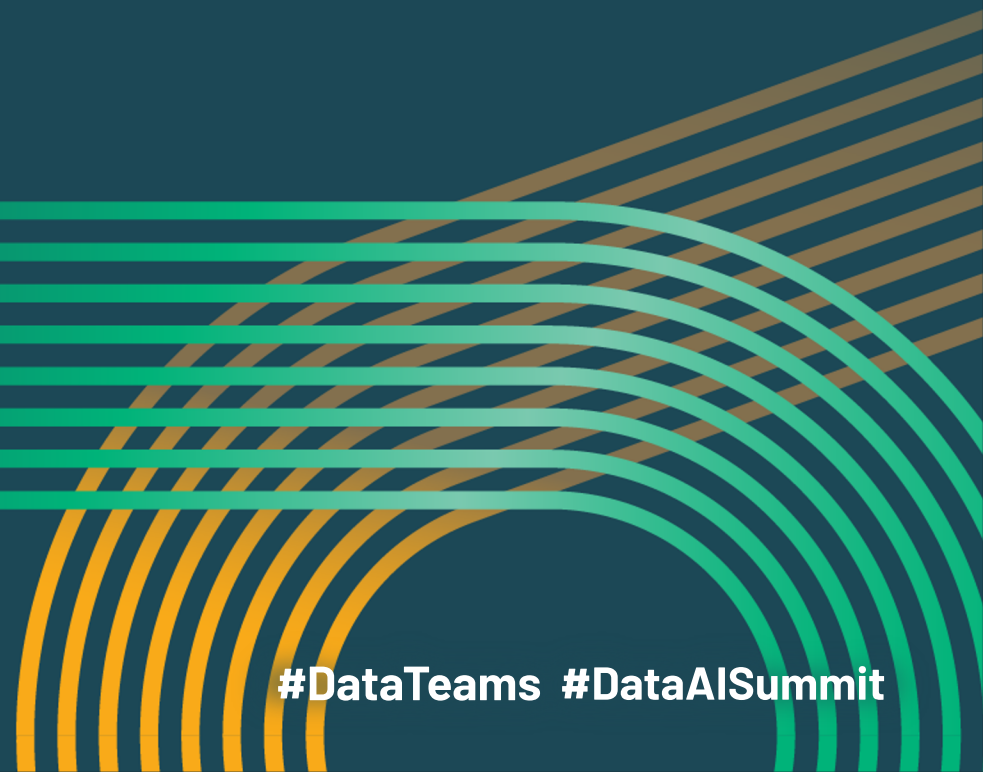
Agenda

- Text Extraction Overview
- **T**ext **D**etection(TD)
- **T**ext **R**ecognition(TR) training data preparation
- CRNN-CTC model for TR
- Attention – OCR
- Spatial Transformer Nets for improving TR accuracy
- Model Accuracies on different dataset.
- Training & Deployment.
- Questions ?

Text Extraction Overview

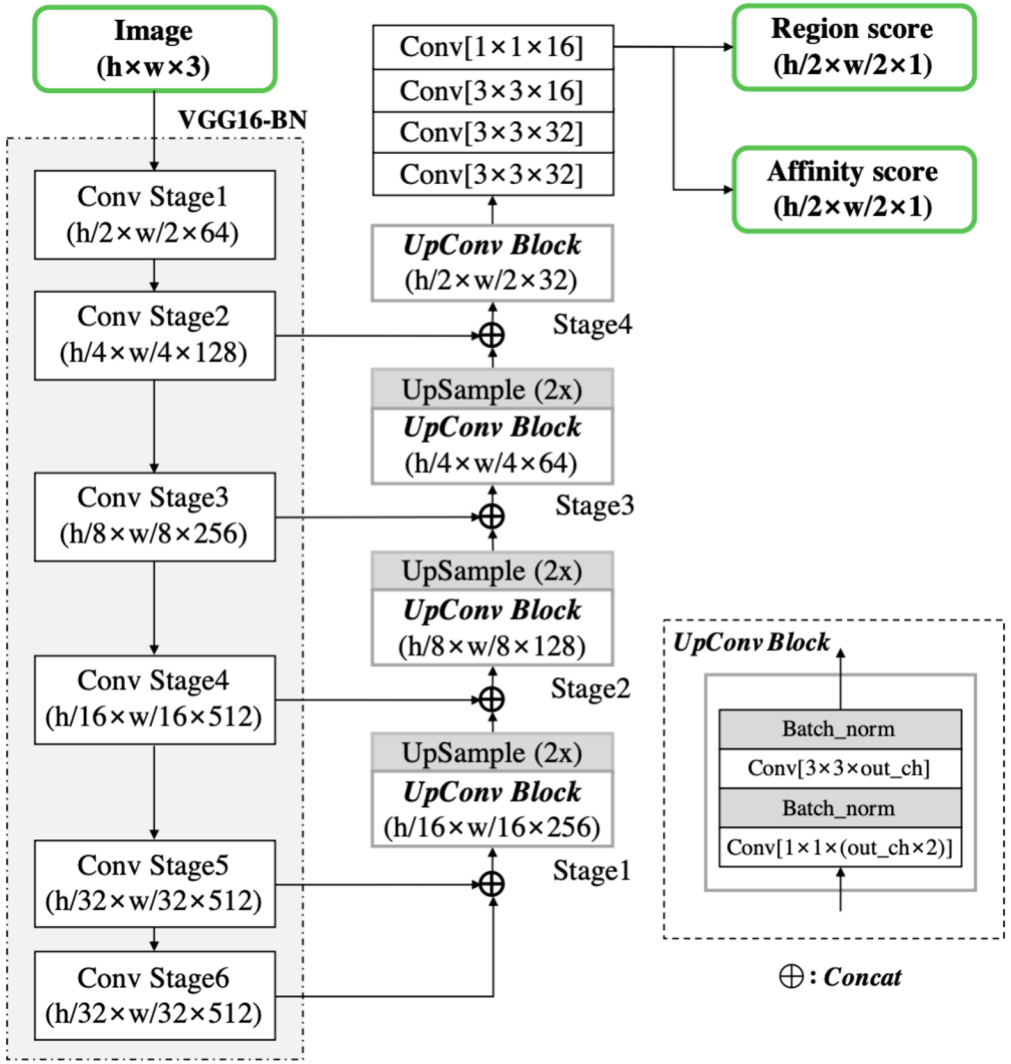


Text Detection



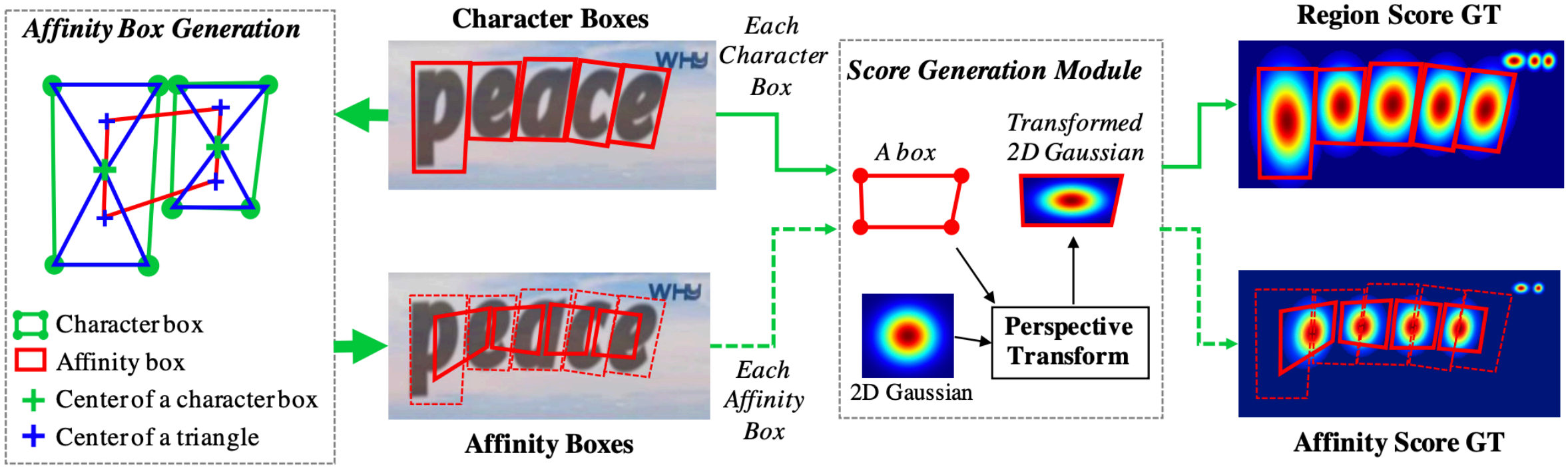
Text Detection – Model architecture

- VGG16 – BN as the backbone
- Model has skip connection in decoder part which is similar to U-Nets.
- Output :
  - Region score
  - Affinity score - grouping characters



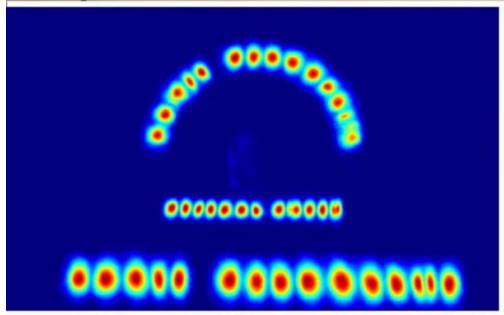
Ref: *Baek, Youngmin, et al. "**C**haracter **R**egion **A**wareness for **T**ext detection." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.*

Ground Truth Label Generation

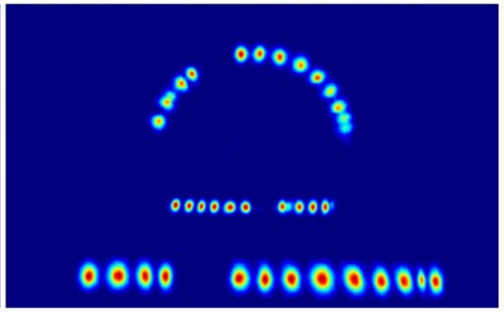


Ref:Baek, Youngmin, et al. "Character region awareness for text detection." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.

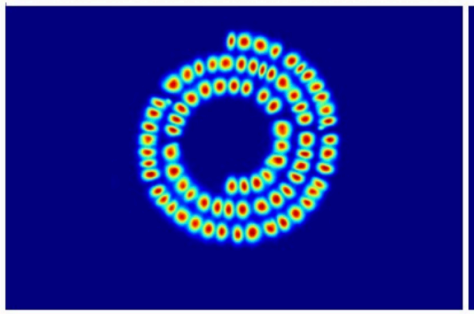
Sample Output



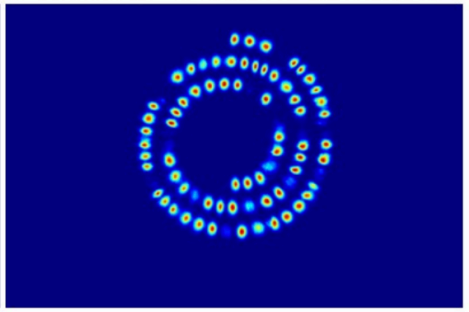
Region Score



Affinity Score



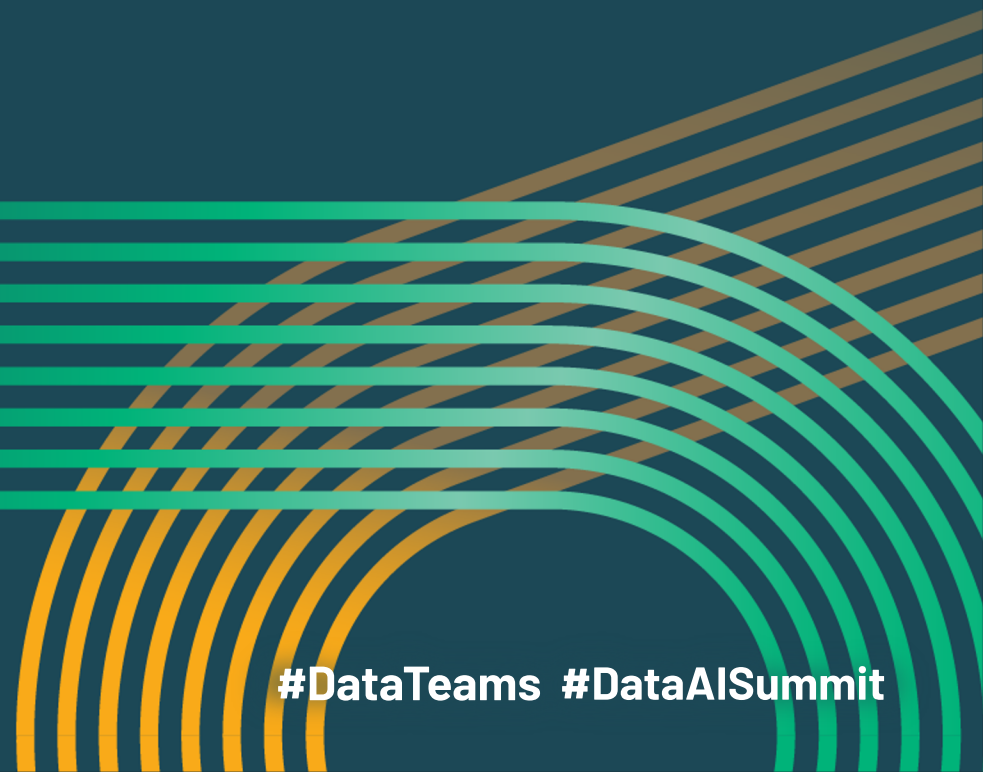
Region Score



Affinity Score

Sample Output ..

Text Recognition



Text Recognition – Training Data Preparation

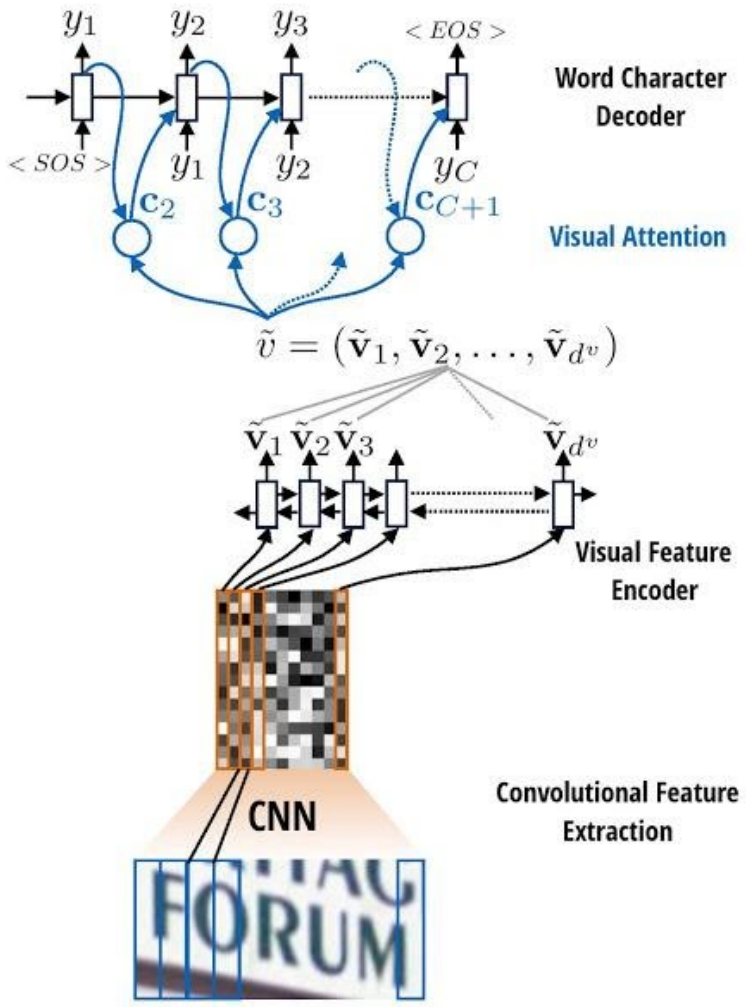
SynthText: image generation engine for building a large annotated dataset.

15 million images generated with different **font styles, size, color** & **varying backgrounds** using product descriptions + open source datasets

Vocabulary: 92 characters
Includes capital + small letters, numbers and special symbols

Link to “Text Recognition with **CRNN-CTC** model” blog published in WANDB : https://bit.ly/3hBaWQv

Attention - OCR

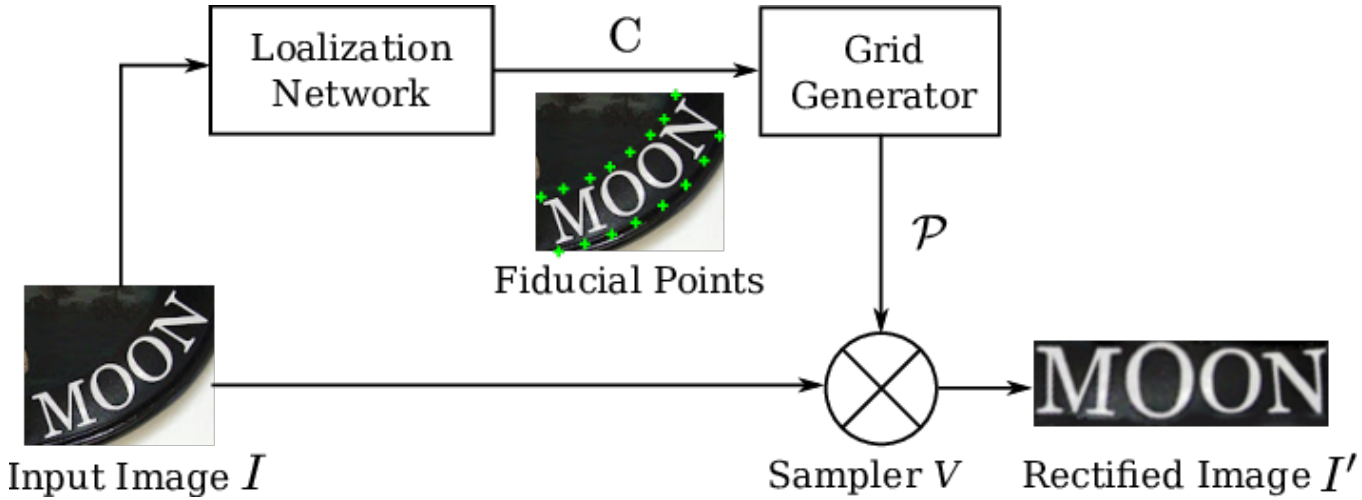


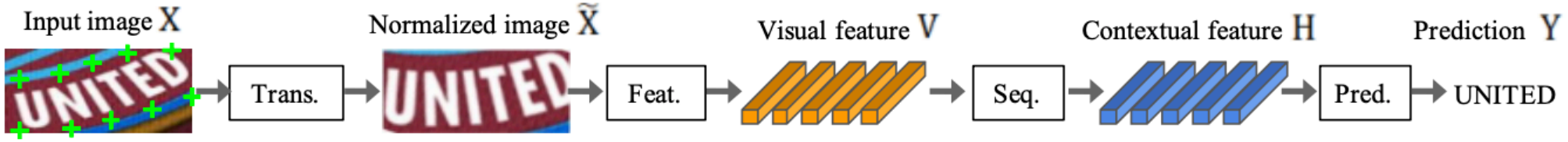
- Encoder – Decoder framework
- CNN used as visual feature encoder.
- LSTM with Attention mechanism is used to extract text in a generative fashion.
- Cross-entropy as a loss function

Product Images with curved text

Spatial Transformation Networks

- Spatial Transformer Network is a learnable module aimed at increasing the spatial invariance of Convolutional Neural Networks in a computationally and parameter efficient manner.





Model Accuracy on Regular and Arbitrary shaped text

Dataset	CRNN-CTC	CNN-LSTM-Attn	STN-CRNN-CTC	STN-CNN-LSTM- Attn
IIIT 5K	81.6	82.1	85	85.16
SVT	82.9	83.5	88.7	88.8
ICDAR03_860	89.2	89.8	91.03	91.7
ICDAR03_867	91.1	91.0	91.59	92.4
ICDAR13_857	92.6	92.7	93.08	94.00
ICDAR13_1015	93.1	93.1	93.25	94.53
ICDAR15_1811	69.4	69.8	72.3	76.5
ICDAR15_2077	64.2	64.8	67.5	71.89
SVT-P	70	70.6	69.4	76.89
CUTE	65.5	66.7	85.7	83.3

Accuracy

Ground truth: Hello
Predicted: Hello } 1

Ground truth: Hello
Predicted: Helo } 0

Dataset mainly with arbitrary shaped text

Training and deployment

- **15 million images ~ 690 GB when loaded into memory!!** Given that on an average images are of the shape **(128 * 32 * 3)** and **dtype is float32.**
- Usage Generators to load only single batch in memory.
- Deployed on Machine Learning Platform internal to Walmart.
- Both text detection and recognition are deployed on single V100 GPU's and prediction time is ~0.45 seconds for each image.

The Team behind the project

Rajesh Shreedhar Bhat
Senior Data Scientist

Pranay Dugar
Data Scientist

Anirban Chatterjee
Staff Data Scientist

Vijay Agneeswaran
Director- Data Science

Sample Code + PPT

https://github.com/rajesh-bhat/data-ai-summit-2020

Questions ??

rsbhat@asu.edu

https://www.linkedin.com/in/rajeshshreedhar