LineEX: Data Extraction from Scientific Line Charts

Shivasankaran V P*; Muhammad Yusuf Hassan*; Mayank Singh

IIT Gandhinagar

vp.shivasan@iitgn.ac.in, md.hassan@iitgn.ac.in, singh.mayank@iitgn.ac.in

1. Motivation

A core problem in the machine readability of scientific line charts: recovering the underlying data present in these charts.

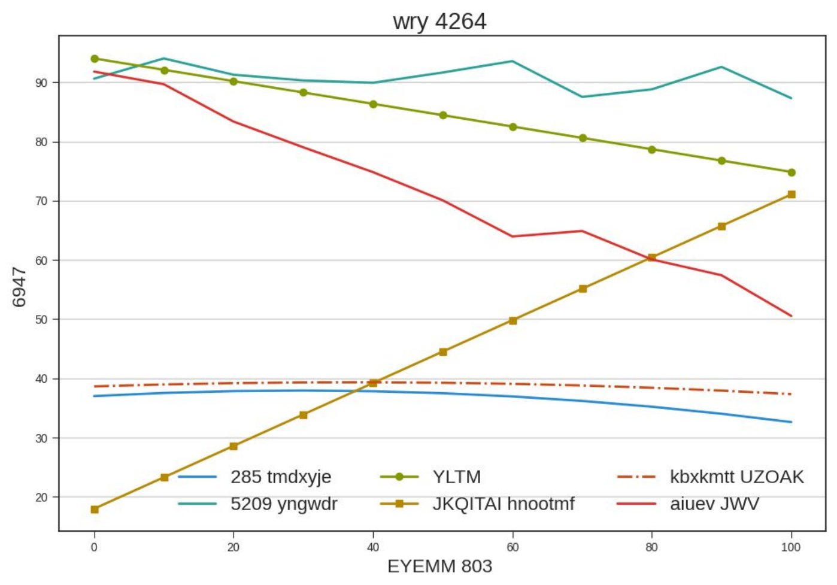
Chart data extraction can be used in downstream tasks like search engines, machine-generated leaderboards, and assistants for impaired people.

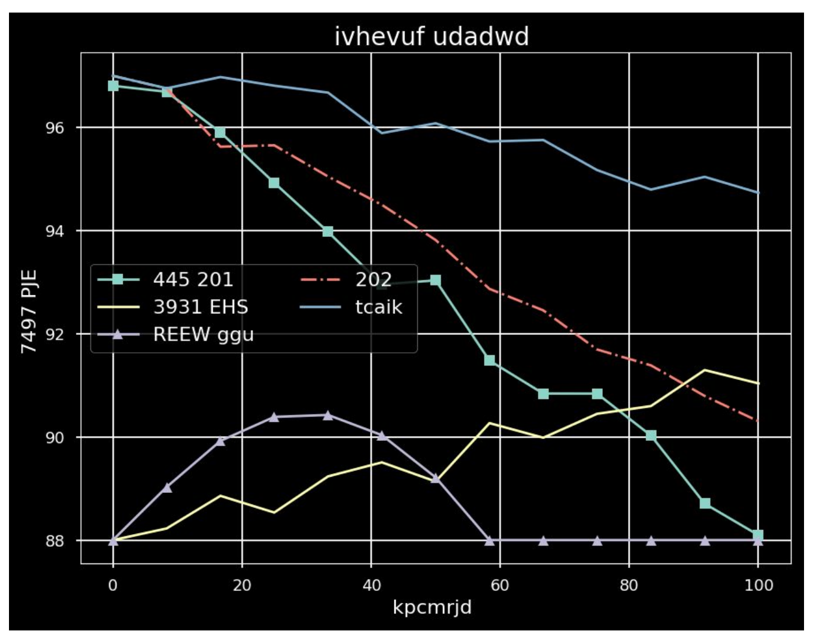
Alphabet Inc (GOOG) PE Ratio (Forward 1y)
Apple Inc (AAPL) PE Ratio (Forward 1y)
Hi: 28.37
Hi: 28.23
28.00
24.60 AVG
21.97 AVG
20.20
16.19
Low: 16.19
Oct '21
Jan '22
Apr '22
Jul '22
Seeking Alpha
Oct 03 2022, 6:13AM EDT. Powered by YCHARTS

2. Dataset

We introduce the largest publicly available dataset of line charts, comprising 430k samples. We make variation in the following aspects:

(i) no. of lines (2-6) (ii) line color (iii) marker style (iv) gridlines (v) plotting style (choice of 27)

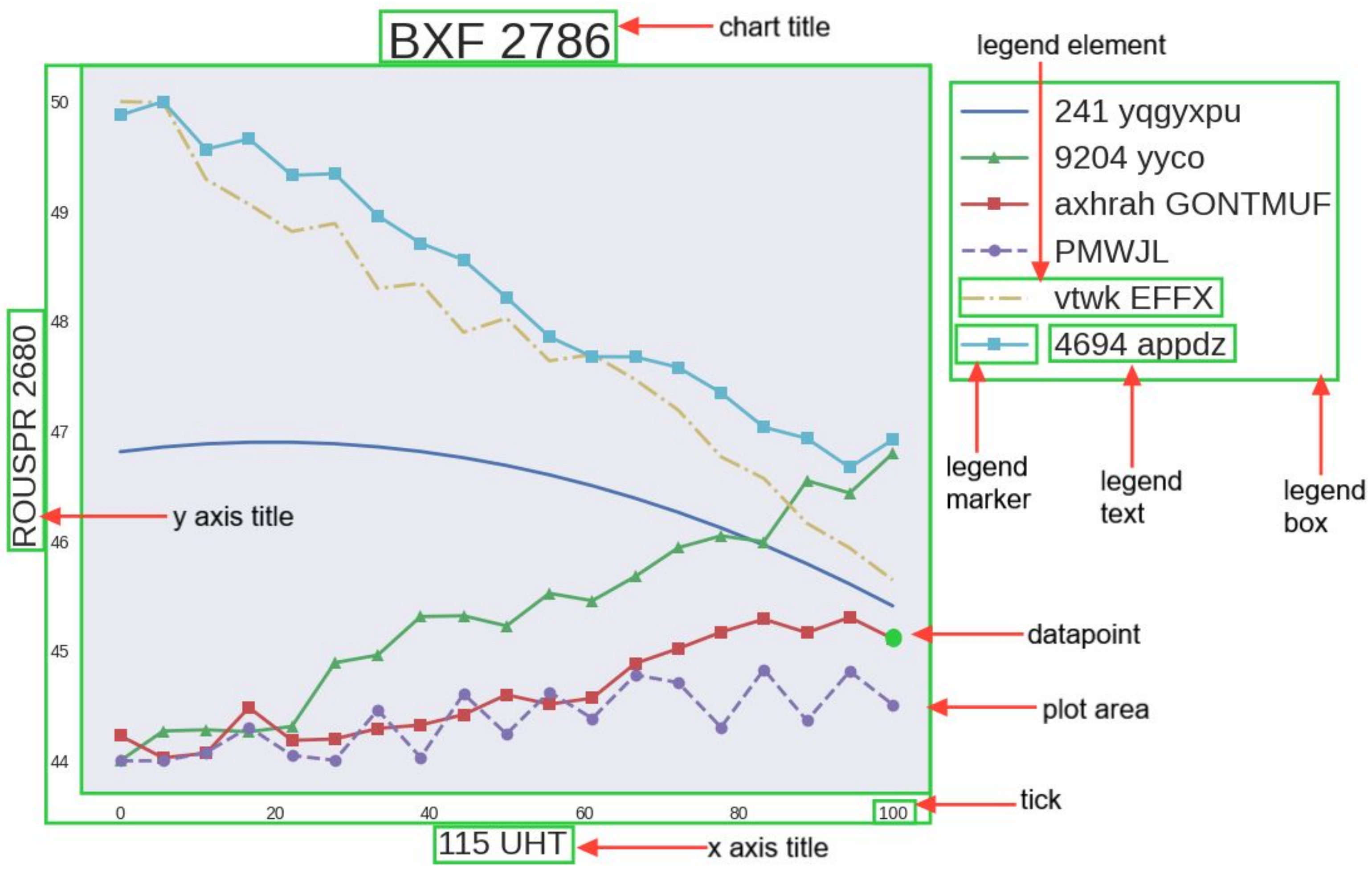




3. Chart Element Detection

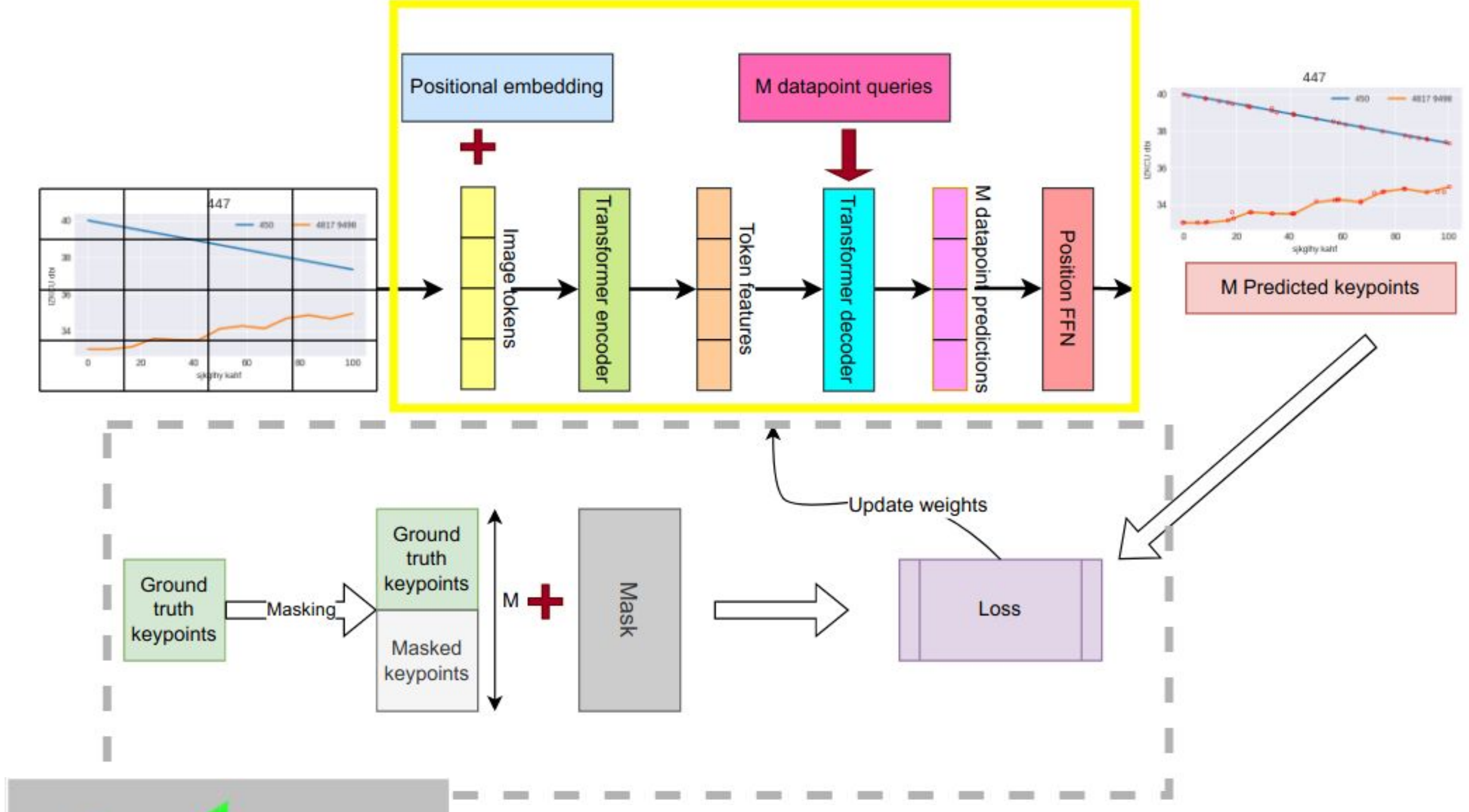
This module is essential for chart understanding and downstream tasks like data point scaling and legend mapping. It consists of two submodules:

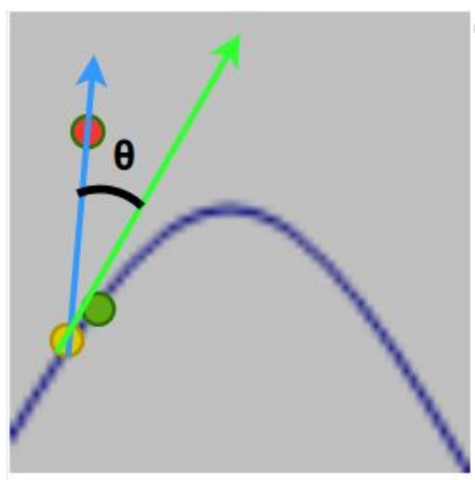
Detection: Detects and localizes elements important for chart understanding. We leverage the DETR (Carion et al.) transformer architecture for this task.
Text Extraction: Extracts the chart's title text, axes labels, and legend texts using the EasyOCR tool.



4. Keypoint Extraction

We adapt pre-existing keypoint extraction model(P. Panteleris et al.). We also propose a novel loss, *Angular Similarity Error(ASE)*, based on the first-derivative approximate at the ground-truth keypoint. We define an anchor point as a pixel near a ground truth keypoint that lies on the same line as the ground truth keypoint



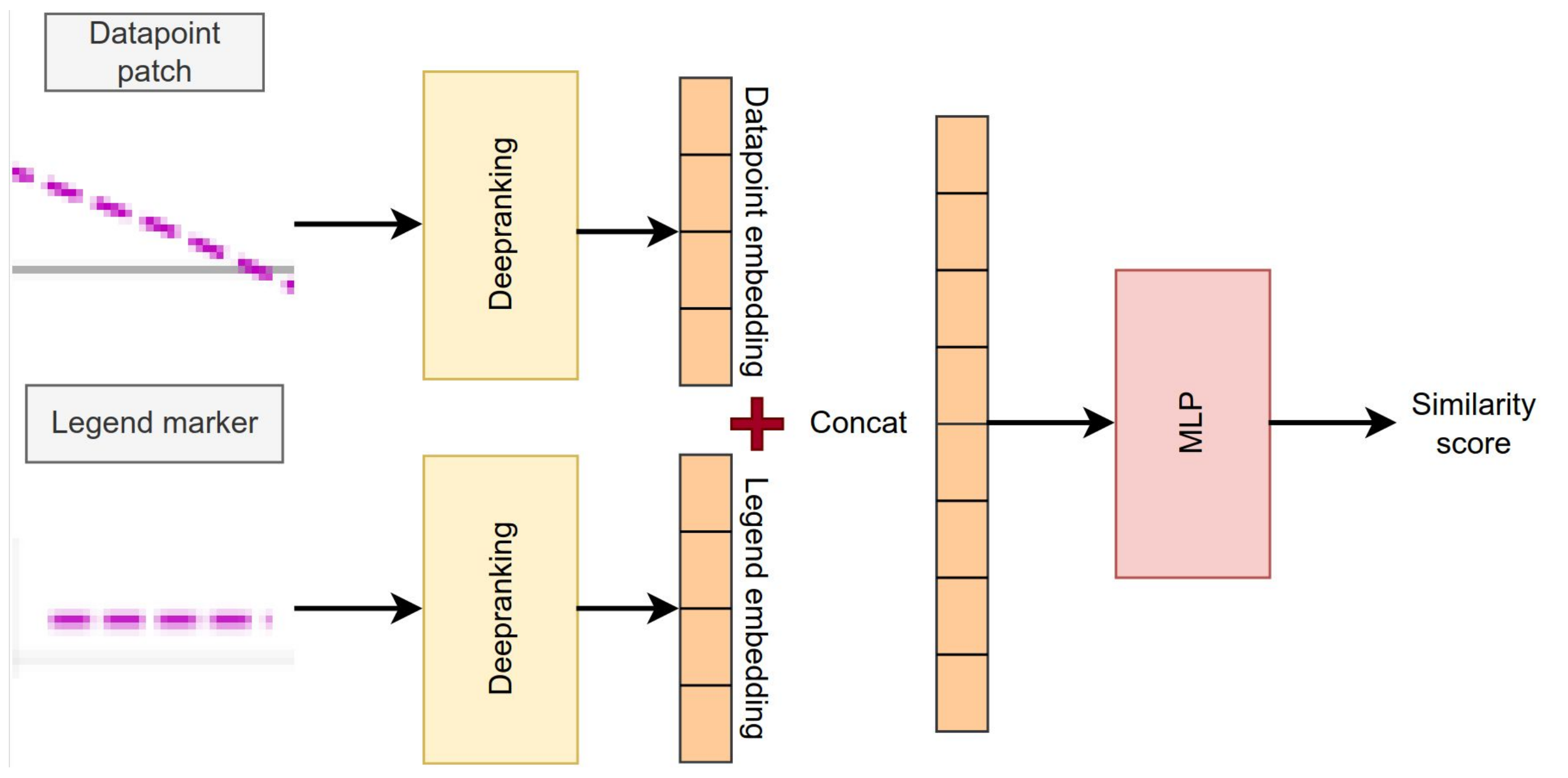


The green arrow is a vector that joins the anchor point and the ground truth keypoint. The Blue arrow is a vector joining the ground truth and the predicted keypoints. ASE minimizes the angle θ.

4. Postprocessing

Legend mapping and line grouping are done based on the similarity between legend and keypoint patches leveraging the DeepRanking (Wang et al.) model.

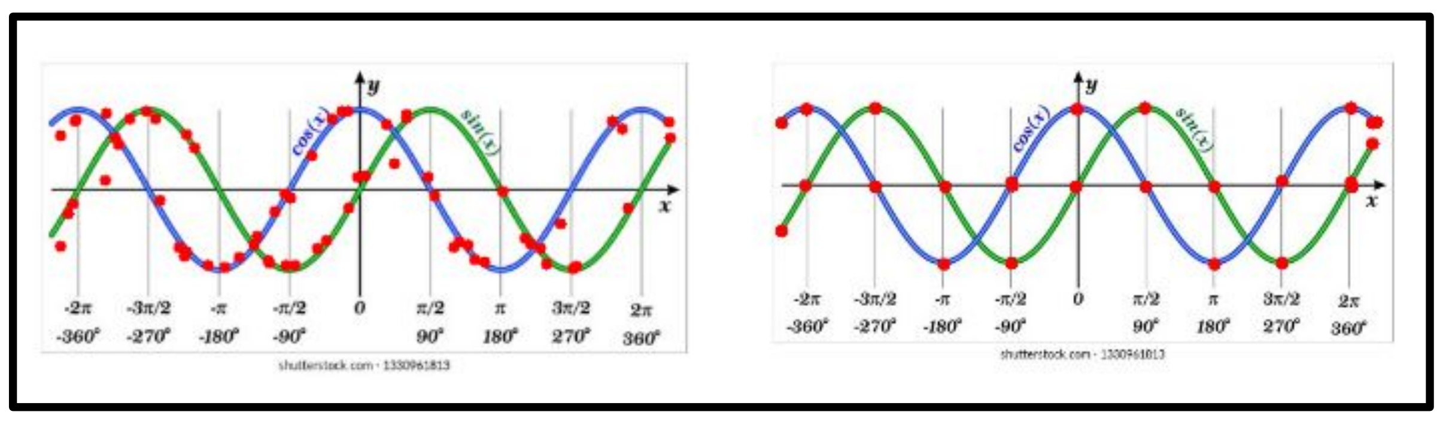
Keypoints are scaled from pixel coordinates to raw data points using ticks information extracted in section 3.



5. Results

Our approach shows significant improvement from previous SOTA.

		ExcelChart400K			Adobe Synthetic			Ours			Smooth		
		Recall	Prec	F1	Recall	Prec	F1	Recall	Prec	F1	Recall	Prec	F1
sim_{str}	ChartOCR	**0.85**	**0.98**	**0.90**	0.76	**0.72**	0.71	0.71	**0.90**	0.78	0.36	**0.74**	0.46
	LINEEX$_D$	0.82	0.69	0.70	0.91	0.54	0.64	0.83	0.75	0.76	0.70	0.49	0.56
	LINEEX$_{D+A}$	0.84	0.80	0.78	**0.94**	0.67	**0.74**	**0.86**	0.84	**0.83**	**0.72**	0.52	**0.59**
sim_{rel}	ChartOCR	**0.85**	**0.98**	**0.90**	0.78	0.80	0.76	0.74	**0.97**	0.83	0.38	**0.78**	0.49
	LINEEX$_D$	0.83	0.87	0.83	0.93	0.76	0.81	0.85	0.92	0.87	0.75	0.58	0.64
	LINEEX$_{D+A}$	0.85	0.90	0.85	**0.93**	**0.81**	**0.84**	**0.87**	0.94	**0.89**	**0.77**	0.61	**0.67**



Left - LineEX; Right ChartOCR