Reading and Reasoning with Text In Images

Ernest Valveny Llobet (Ernest.Valveny@uab.cat)

Centre de Visió per Computador

Outline

- Datasets for Text Detection and Recognition
- Text Detection
  - General detection-based methods
  - Sub-text components methods
- Text Recognition
  - Classification-based methods
  - Sequence-based methods
- End-to-end Reading systems
- Current challenges of Reading systems
  - Commercial OCR systems
- Reasoning with Text
  - Visual Question Answering

Text Understanding in the Wild

S.R. Battu, M. Mathew, L. Gomez, M. Russinyol, D. Karatzas, C.V. Jawahar, "RoadText-1K : A Dataset for Text Detection and Recognition in Driving Videos", ICRA 2020

R. Gomez, A. Biten, L. Gomez, J. Gibert, D. Karatzas, M. Rusiñol. "Selective style transfer for text". ICDAR 2019

L. Gómez, A. Mafla, M. Rusinol, D. Karatzas. "Single Shot Scene Text Retrieval". ECCV 2018

L. Gómez, & D. Karatzas, "Textproposals: a text-specific selective search algorithm for word spotting in the wild". Pattern Recognition, 2017

L. Gomez, et al, "Improving patch-based scene text script identification with ensembles of conjoined networks", Pattern Recognition, 2017

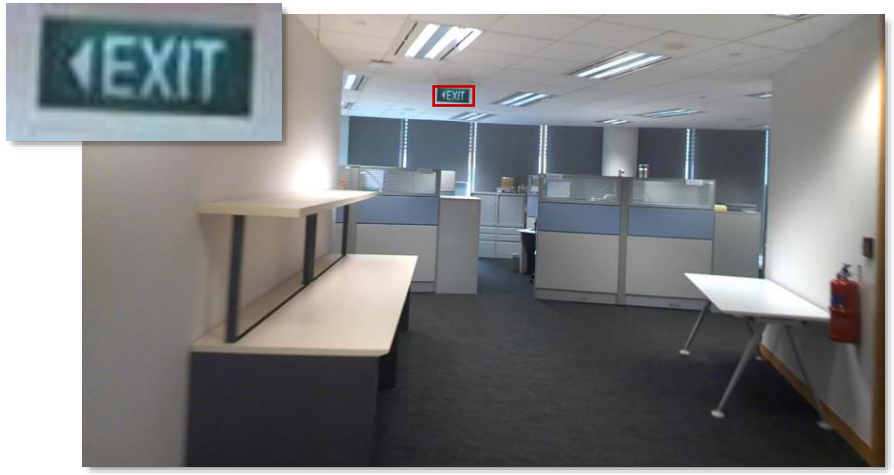
Typical Tasks

- Detection / Localization
- Recognition
- Word spotting
- Information spotting
- Scene text retrieval
- Script identification
- Video text / text tracking
- Text segmentation / replacement
- Text style transfer

Open Challenges

- Video text / text tracking
- Multi-script / multi-lingual scenarios
- Uncontrolled conditions (night, rain, etc)
- etc

Not a solved problem…



Small text instances

Out of vocabulary text

Text in motion

Multiple scripts

Is This Text?

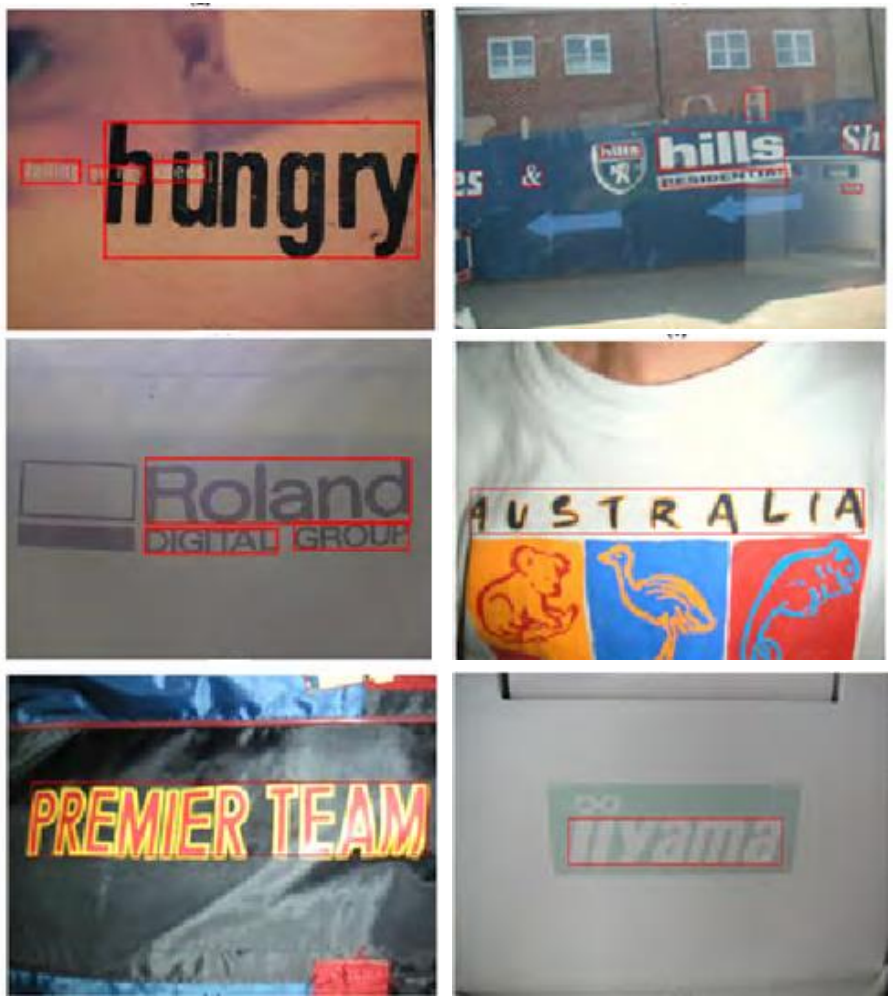
Unlike generic “objects”, as traditionally defined, **text is a composition of non-adjacent parts, is recursively defined**, and its appearance varies drastically, even among different instances with the same granularity (e.g. words).

Lluís Gomez-Bigorda, “Exploiting similarity hierarchies for multi-script scene text understanding”, PhD Thesis, April 2016

Datasets for Scene-Text Detection and Recognition

Datasets - Annotation Types

Horizontal bounding boxes



Karatzas, Dimosthenis, et al. "ICDAR 2013 robust reading competition." 2013 12th International Conference on Document Analysis and Recognition. IEEE, 2013.

Multi-oriented quadrilaterals

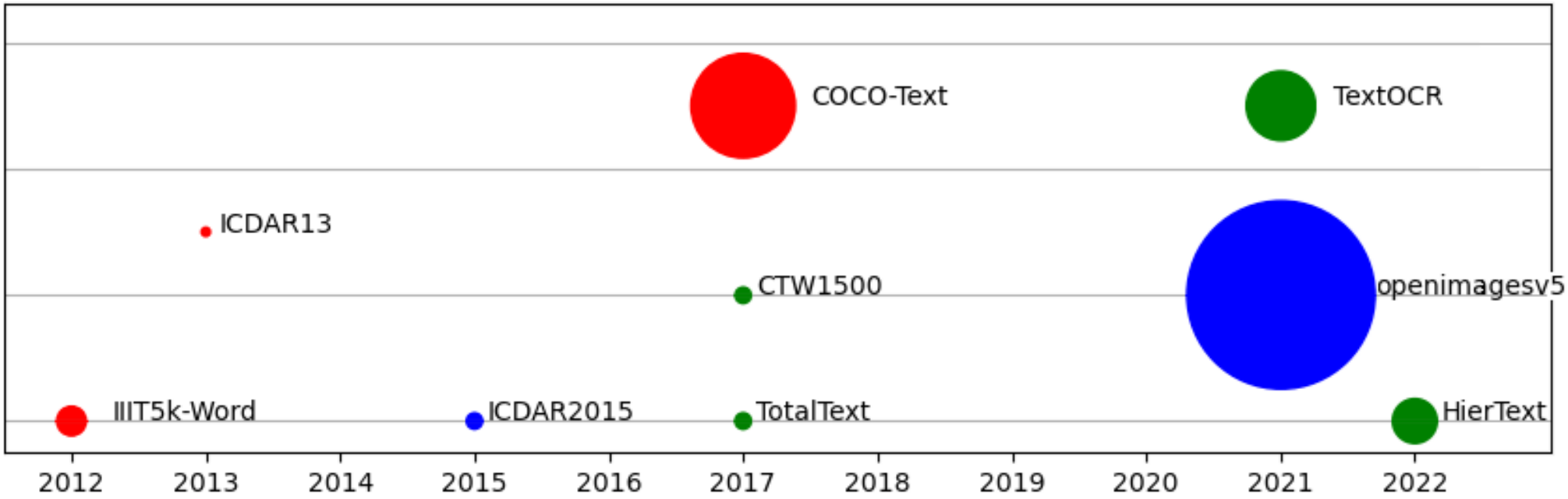
Karatzas, Dimosthenis, et al. "ICDAR 2015 competition on robust reading." 2015 13th international conference on document analysis and recognition (ICDAR). IEEE, 2015.

Datasets - Annotation Types

Polygonal contours

Ch'ng, Chee Kheng, and Chee Seng Chan. "Total-text: A comprehensive dataset for scene text detection and recognition." 2017 14th IAPR international conference on document analysis and recognition (ICDAR). Vol. 1. IEEE, 2017.

Evolution of Dataset Size



Horizontal bounding boxes

Rotated quadrilaterals

Curved (Polygonal)

200k+ Images

462 Images

Karatzas, Dimosthenis, et al. "ICDAR 2013 robust reading competition." *2013 12th International Conference on Document Analysis and Recognition*. IEEE, 2013.

Mishra, Anand, Karteek Alahari, and C. V. Jawahar. "Scene text recognition using higher order language priors." *BMVC-British machine vision conference*. BMVA, 2012.

Karatzas, Dimosthenis, et al. "ICDAR 2015 competition on robust reading." *2015 13th international conference on document analysis and recognition (ICDAR)*. IEEE, 2015.

Veit, Andreas, et al. "Coco-text: Dataset and benchmark for text detection and recognition in natural images." *arXiv preprint arXiv:1601.07140* (2016).

Yuliang, Liu, et al. "Detecting curve text in the wild: New dataset and new solution." *arXiv preprint arXiv:1712.02170* (2017).

Ch'ng, Chee Kheng, and Chee Seng Chan. "Total-text: A comprehensive dataset for scene text detection and recognition." *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*. Vol. 1. IEEE, 2017.

Singh, Amanpreet, et al. "TextOCR: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

Krylov, Ilya, Sergei Nosov, and Vladislav Sovrasov. "Open Images V5 Text Annotation and Yet Another Mask Text Spotter." *Asian Conference on Machine Learning*. PMLR, 2021.

Long, Shangbang, et al. "Towards End-to-End Unified Scene Text Detection and Layout Analysis." *arXiv preprint arXiv:2203.15143* (2022).

Synthetic Data

Gupta 2016

A. Gupta, A. Vedaldi, A. Zisserman Synthetic data for text localisation in natural images, CVPR2016, pp. 2315-2324

Liao 2020 (SynthText3D)

Liao, M., Song, B., Long, S., He, M., Yao, C., & Bai, X. (2020). SynthText3D: synthesizing scene text images from 3D virtual worlds. Science China Information Sciences, 63(2), 1-14.

Long 2020 (UnrealText)

Long, S., & Yao, C. (2020). Unrealtext: Synthesizing realistic scene text images from the unreal world. arXiv preprint arXiv:2003.10608.

Results on some datasets

	Born-digital (ICDAR11)	Focused Text (ICDAR13)	Incidental Text (ICDAR15)	COCO-Text (ICDAR17)	TotalText (2017)	Openimagesv5 (2021)
Best End-to-End Generic vocabulary	86.23	88.73	75.01	43.58	78.7	56.3
						

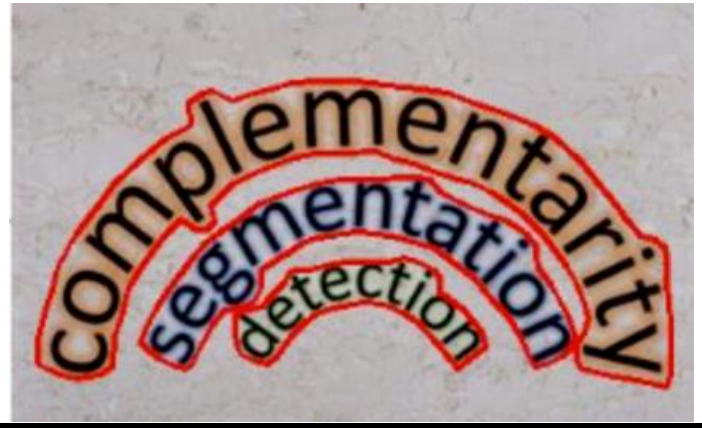
Text Detection

Text Detection

Two main approaches have dominated scene-text detection:

- General detection-based methods: methods based on object detectors such as YOLO, SSD, etc.

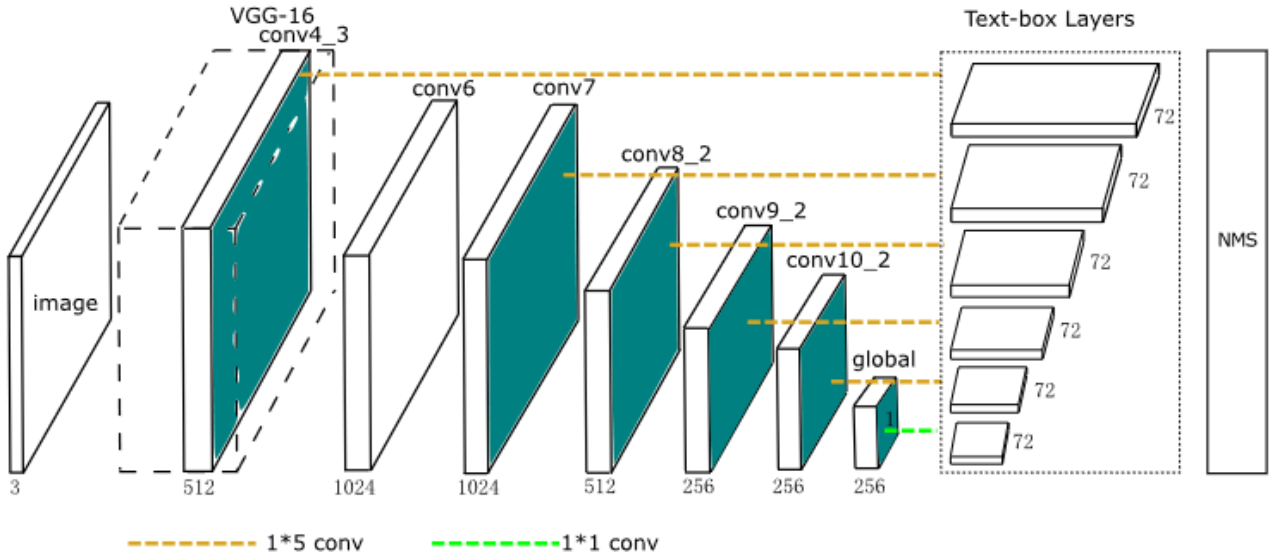
- Segmentation-based methods

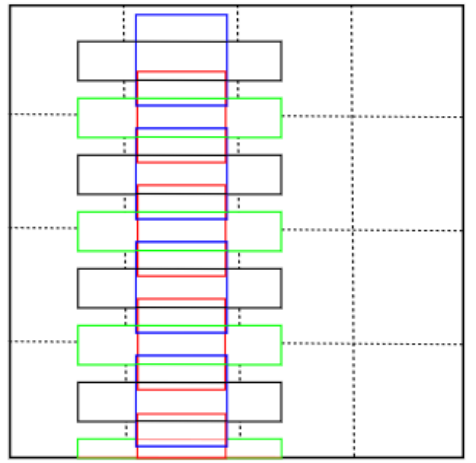


Detection-Based Methods

TextBoxes

- Single-stage detector.
- Fully convolutional network that predicts bounding boxes at different resolutions using a set of specific pre-defined anchor boxes adapted to text aspect ratio
- Outputs are aggregated by NMS





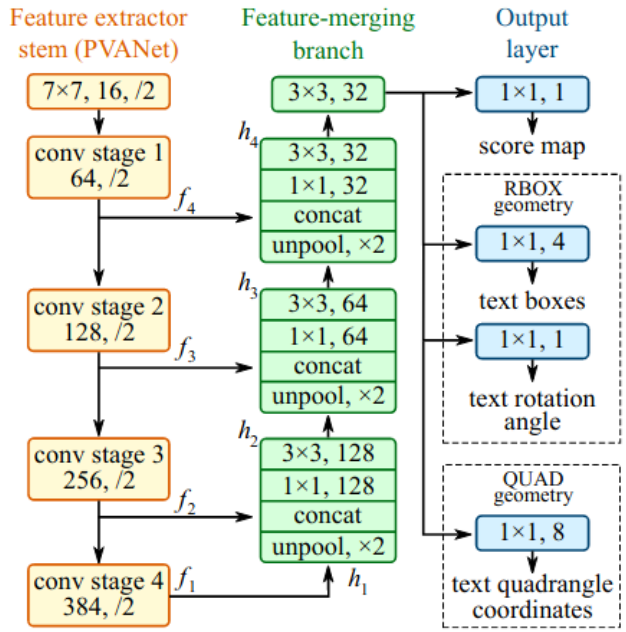
Anchor boxes

Liao, Minghui, et al. (2017) “Textboxes: A fast text detector with a single deep neural network”. Proceedings of the AAAI conference on artificial intelligence. 31:1

Detection-Based Methods

EAST

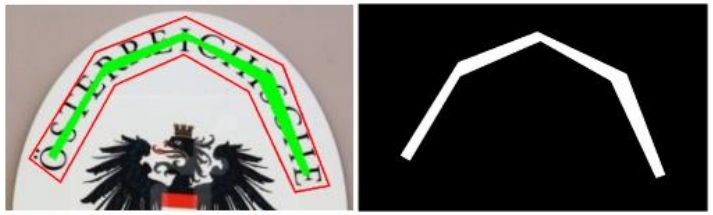
- Single-stage detector inspired in DenseBox detector
- Based on a U-Shape FCN architecture
- No pre-defined anchor boxes
- Predicts rotated bounding boxes
- For every pixel it predicts:
  - Text region score map
  - Geometry of text region (distances to boundaries of the bounding box) + rotation angle
- NMS to obtain the final predictions



Zhou, Xinyu, et al. "East: an efficient and accurate scene text detector." *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2017.

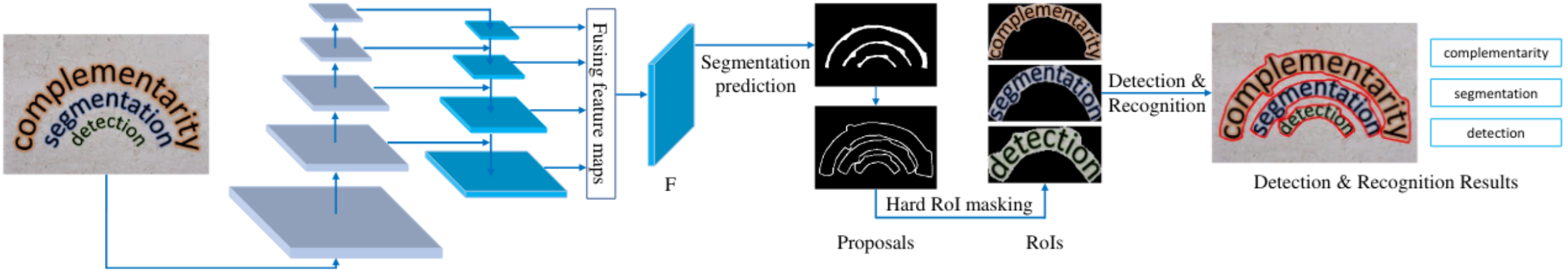
Segmentation-based Methods

Based on obtaining a segmentation map of the text regions that later is refined using some post-processing step.



Mask textspotter v3

- Based on a U-Net segmentatin network
- Generates a segmentation map of the center area of the text region
- The segmentation map is converted into a contour mask that generates a ROI proposal that is refined by a Fast R-CNN network



Liao, Minghui, et al. "Mask textspotter v3: Segmentation proposal network for robust scene text spotting." *European Conference on Computer Vision*. Springer, Cham, 2020.

Text Recognition

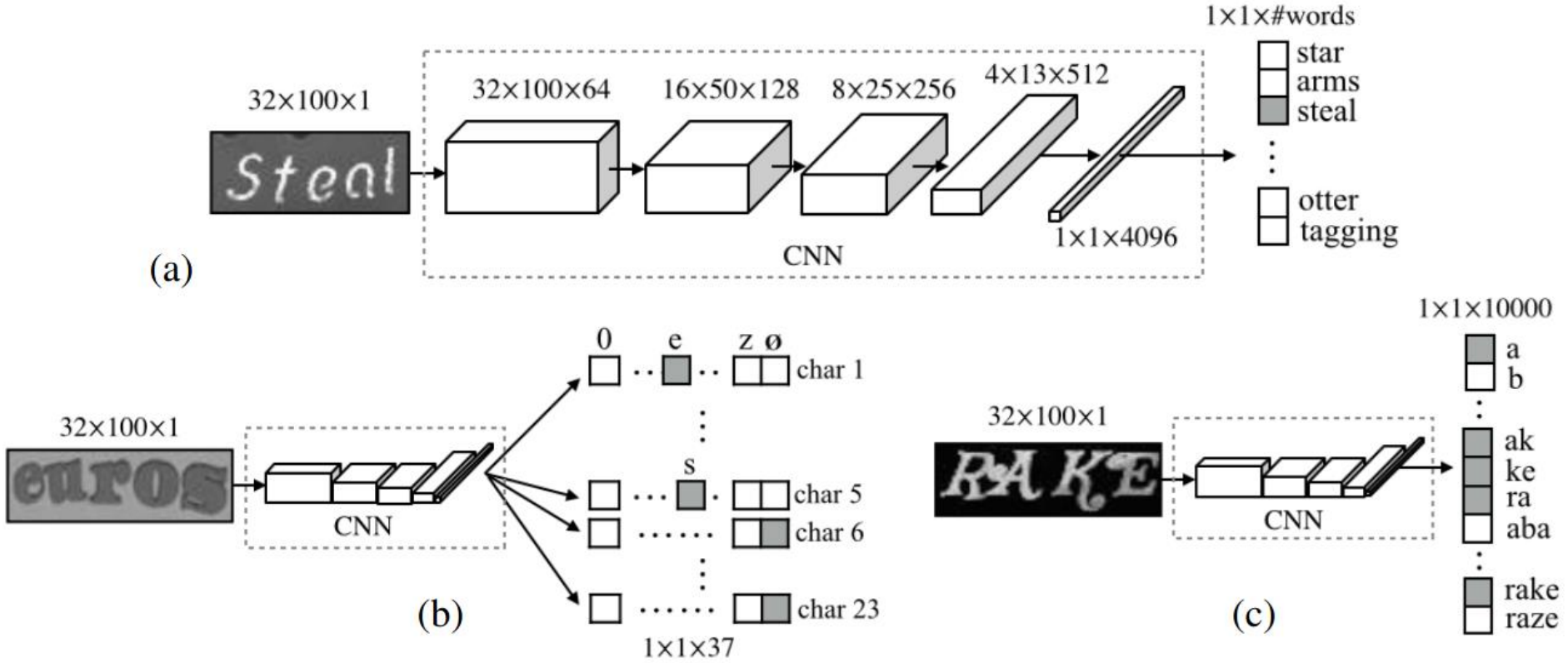
Text Recognition

Text recognition has been approached

- Treating recognition as **a classification problem**: each detection is classified into a class, each class is a word in the dictionary.
- **Sequence-based** methods:
  - **Connectionist temporal classification** (CTC), adopted from speech recognition.
  - **Encoder-decoder networks**, inspired by natural language processing.
- **Transformer-based** methods.

Classification-Based Methods

The first approaches to text recognition treated **recognition as a classification problem**:



Jaderberg, Max, et al. "Synthetic data and artificial neural networks for natural scene text recognition." arXiv preprint arXiv:1406.2227 (2014).

Sequence-Based Methods

The most common approach to text recognition involves the following steps:

- **Text rectification**
- **1D feature slices** (feature extraction)
- **Sequential recognition** (CTC/encoder-decoder)

Different rectification methods have been proposed, such as **region rotation** (for example RoI-rotate from FOTS) or **region unwarping** (for example BezierAlign, from ABCNet).

Liu, Xuebo, et al. "Fots: Fast oriented text spotting with a unified network." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

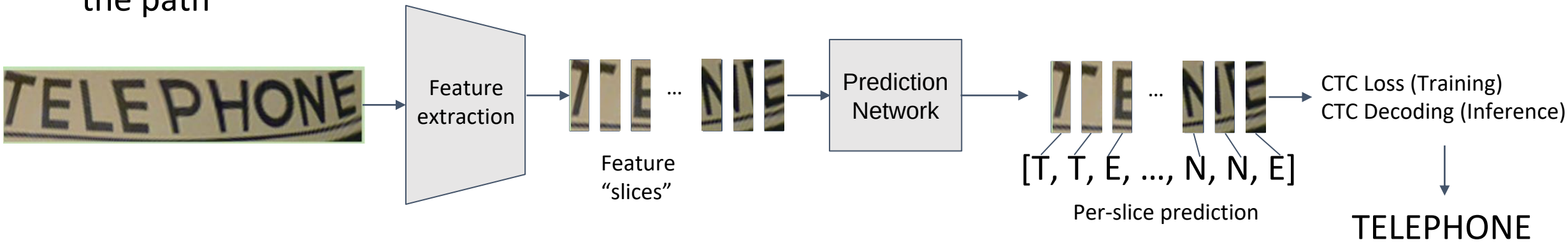
Liu, Yuliang, et al. "Abcnet: Real-time scene text spotting with adaptive bezier-curve network." proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

Connectionist Temporal Classification

Connectionist Temporal Classification (CTC) is a technique that comes from speech recognition.

CTC-Based methods generate **one prediction per-slice**, and use CTC-decoding to produce the final prediction:

- **Training**: CTC loss sums over the probability of all possible alignments of per-slice predictions to the ground-truth.
- **Inference**: CTC Decoding calculates the best path by taking the most likely per-slice predictions and obtains the recognized text by removing duplicate characters and removing all blanks from the path



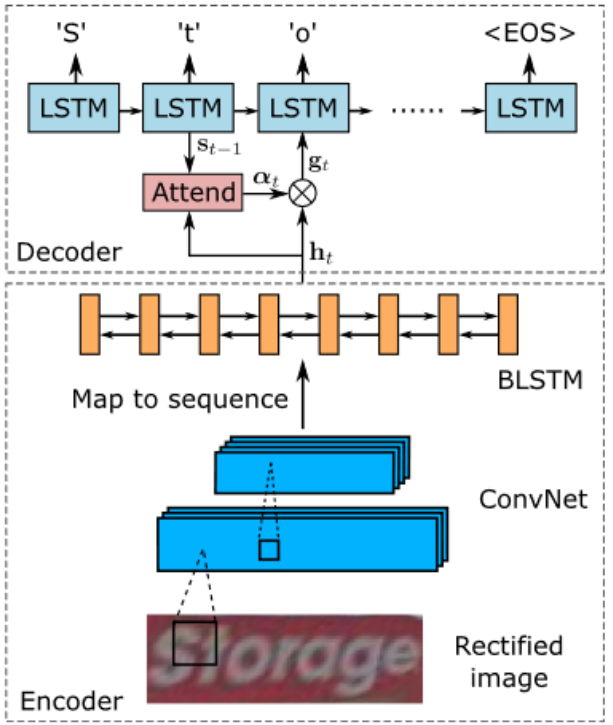
Graves, Alex, et al. "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks." Proceedings of the 23rd international conference on Machine learning. 2006.

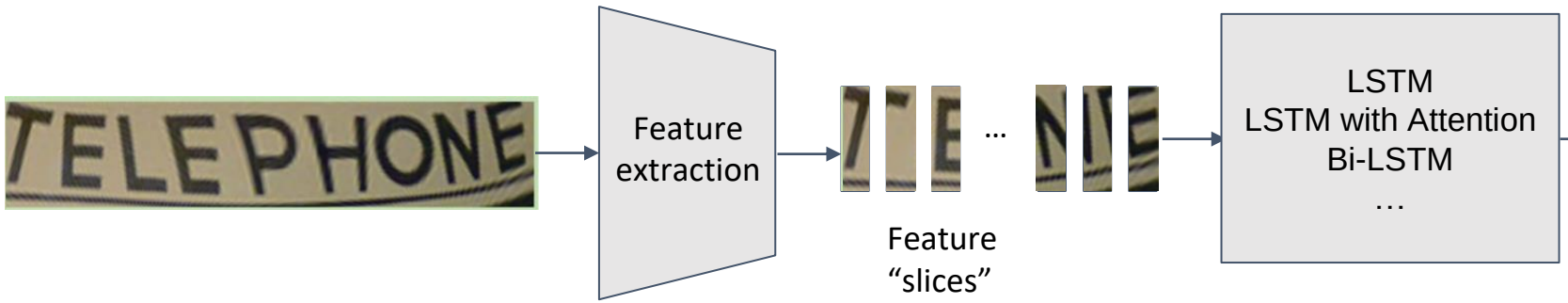
Encoder/Decoder Networks

Encoder-Decoder based recognition uses methods that come from natural language processing (NLP) such as LSTMs, bidirectional LSTMs, etc.

The encoder-decoder network predicts a **variable length** output, and usually uses cross-entropy as the loss of the model.

Example: ASTER model

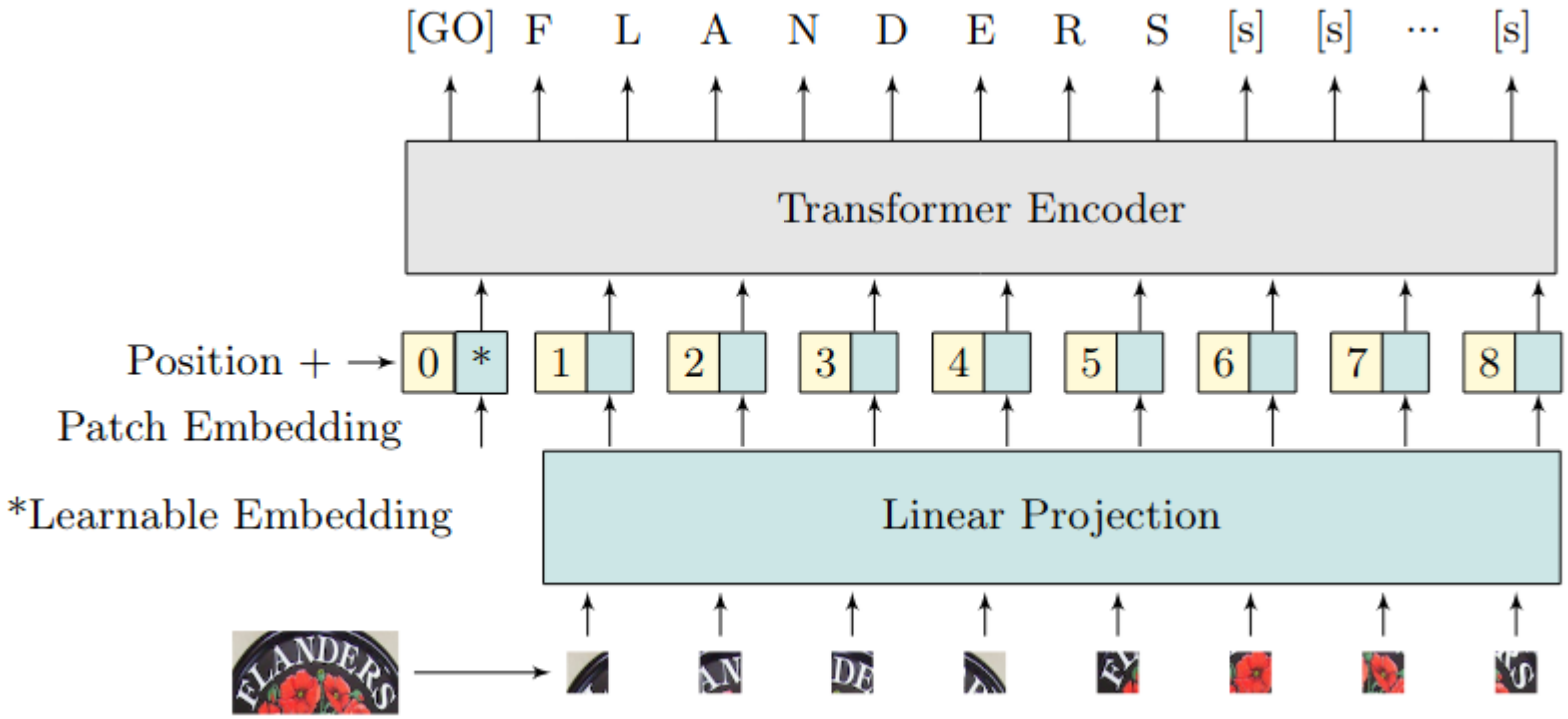




B. Shi et al. "ASTER: An Attentional Scene Text Recognizer with Flexible Rectification" *IEEE Transactions on Pattern Analysis and Machine Intelligence. 2019*

Transformer-Based

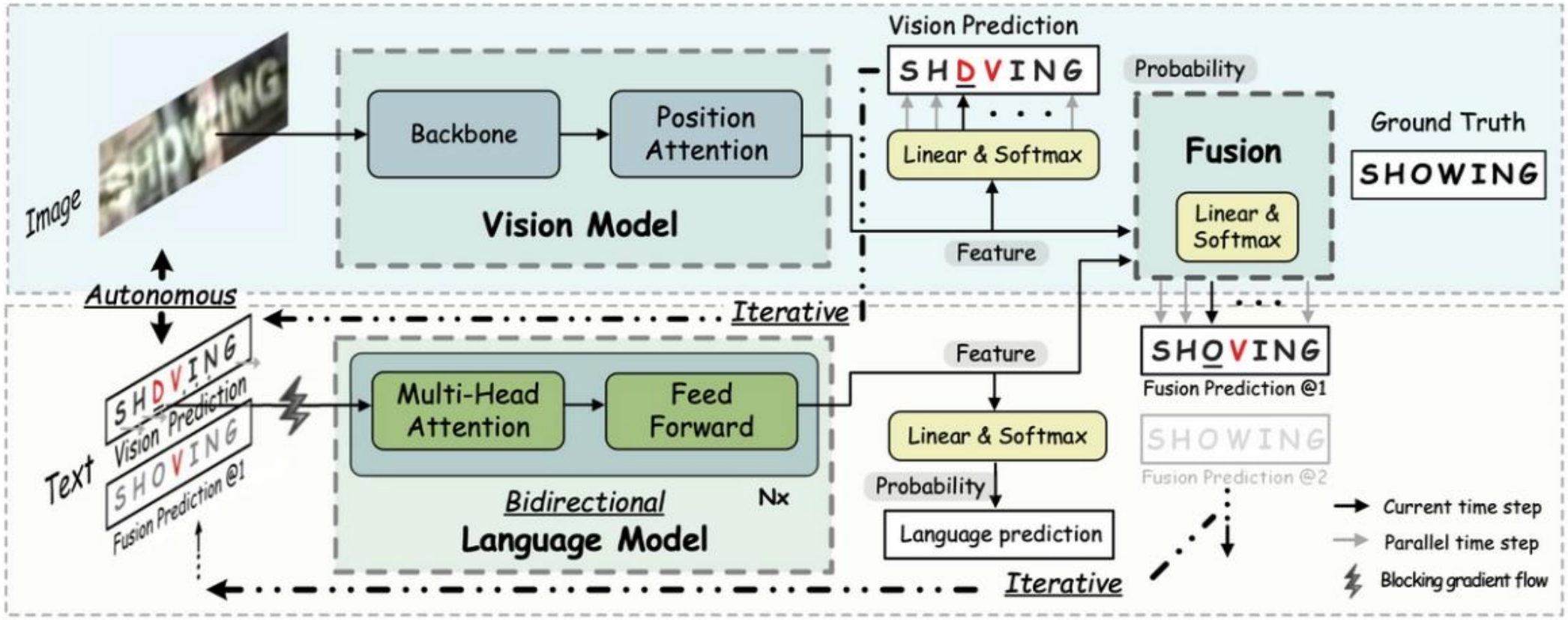
Transformers have started being used for text recognition. ViTSTR uses a **ViT-based transformer** to recognise the word into learnable tokens.



Atienza, Rowel. "Vision transformer for fast and efficient scene text recognition." International Conference on Document Analysis and Recognition. Springer, Cham, 2021.

Recovering From Errors With Language

ABInet uses a language model to overcome OCR errors when



Fang, Shancheng, et al. "Read like humans: autonomous, bidirectional and iterative language modeling for scene text recognition." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

End-to-End Text Detection and Recognition

End-To-End Systems

End-To-End systems combine detection and recognition networks to perform detection and recognition, there are several approaches:

- The usual approach is to have **two steps**, the first is the **text detection** and the second is the **recognition**. In these we make two distinctions:
  - **Two-step pipelines** that are not trainable end-to-end.
  - **Two-stage pipelines** that can be trained end-to-end.
- More recently, some models have started adopting **one-stage models**, sometimes using transformers.

Two Step Models

Text Detection

-Bounding boxes, contours, segmentation, etc.

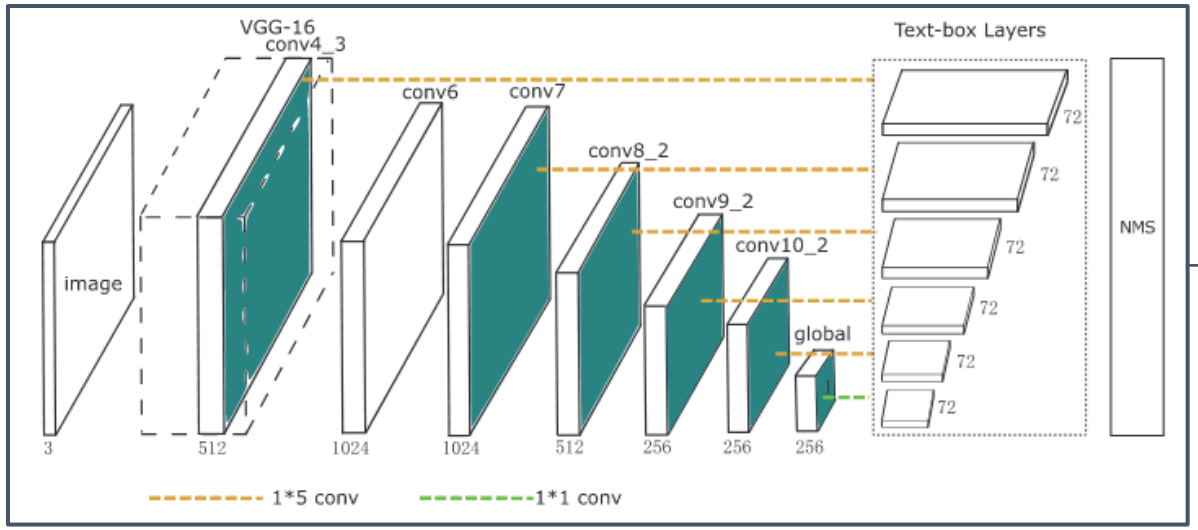
Text Recognition

BANDON

PRESERVE

End-To-End Systems

Textboxes is an example of a two stage framework that is **not trainable end-to-end**:



SSD-based detector

Detection proposals

Crop

RCNN+CTC

"Times"

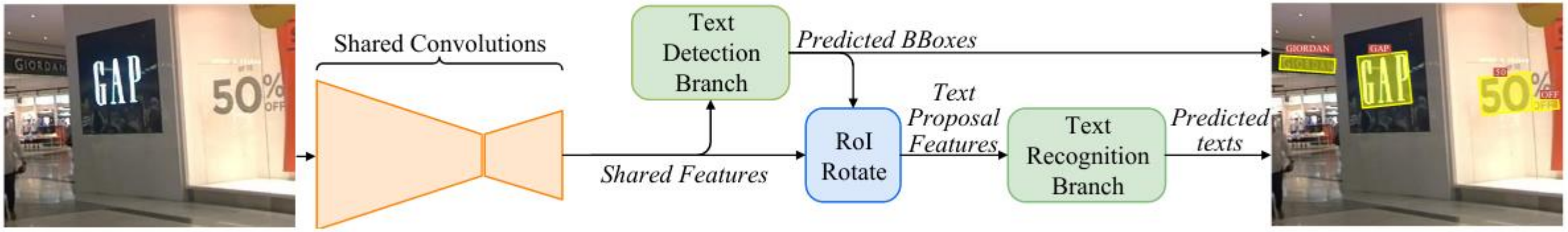
In these approaches there's a **lack of gradient flow** between the detection and the recognition network, and **the features are not shared**.

Liao, Minghui, et al. "Textboxes: A fast text detector with a single deep neural network." *Thirty-first AAAI conference on artificial intelligence*. 2017.

End-To-End Systems

The previous models were followed by **end-to-end trainable networks**. They have the advantage of sharing features and having common gradient flow.

FOTS (Fast Oriented Text Spotting) is an example of such networks. The recognition branch uses **crops of the output feature map**.

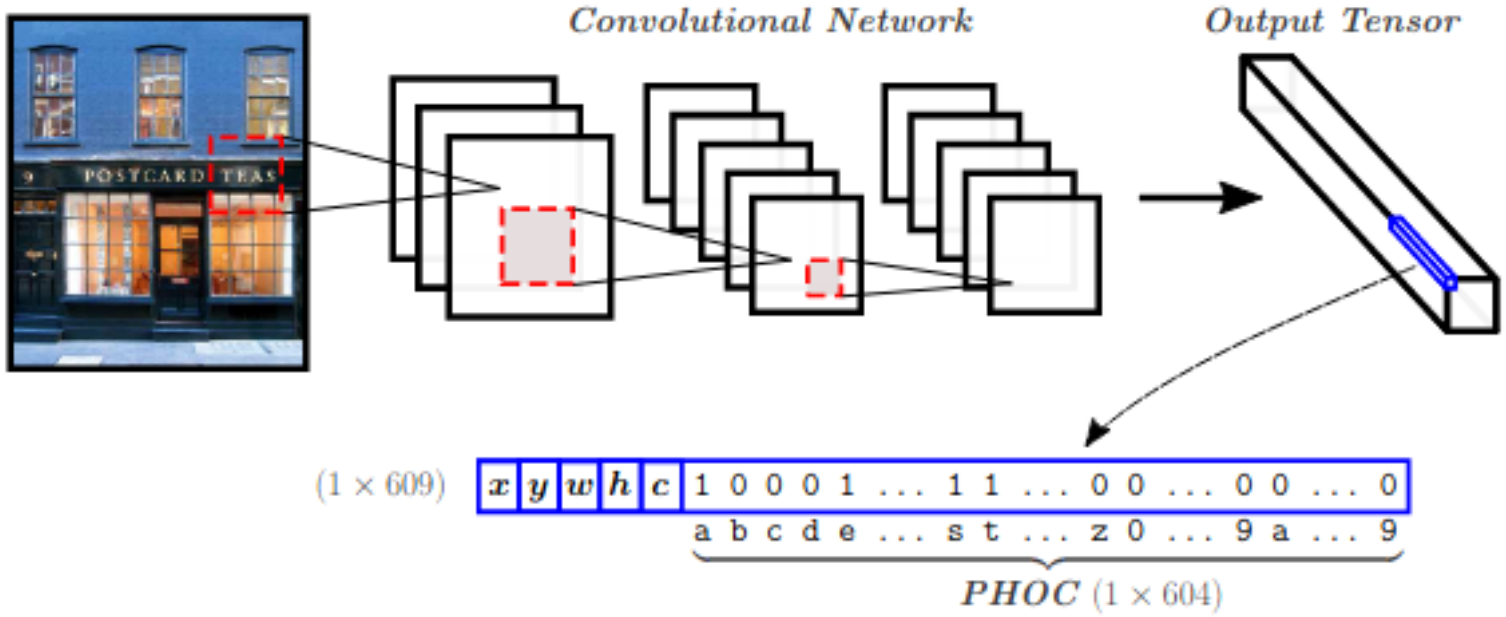


FOTS rotates the cropped feature maps before passing it to the CTC-based decoder.

Liu, Xuebo, et al. "Fots: Fast oriented text spotting with a unified network." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

End-To-End Systems

One of the first **end-to-end trainable**, **one stage** methods was the YOLO-PHOC. This model was used for text retrieval and predicts the PHOC representation of each detected object.

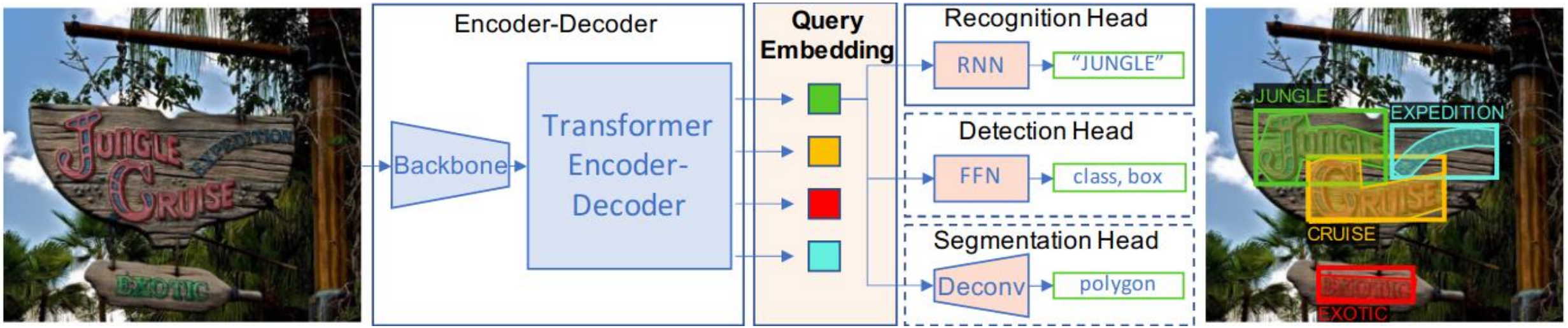


Gómez, Lluís, et al. "Single shot scene text retrieval." *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.

End-To-End Systems

More recent advancements propose **single stage** architectures that don't require **ROI de-warping operations**.

For example, TextTranSpotter uses a encoder-decoder transformer. This model can work with **different types of annotations**, such as polygonal contours or horizontal boxes.



Kittenplon, Yair, et al. "Towards Weakly-Supervised Text Spotting using a Multi-Task Transformer." *arXiv preprint arXiv:2202.05508* (2022).

Current Challenges

OCR Systems

ABBYY OCR

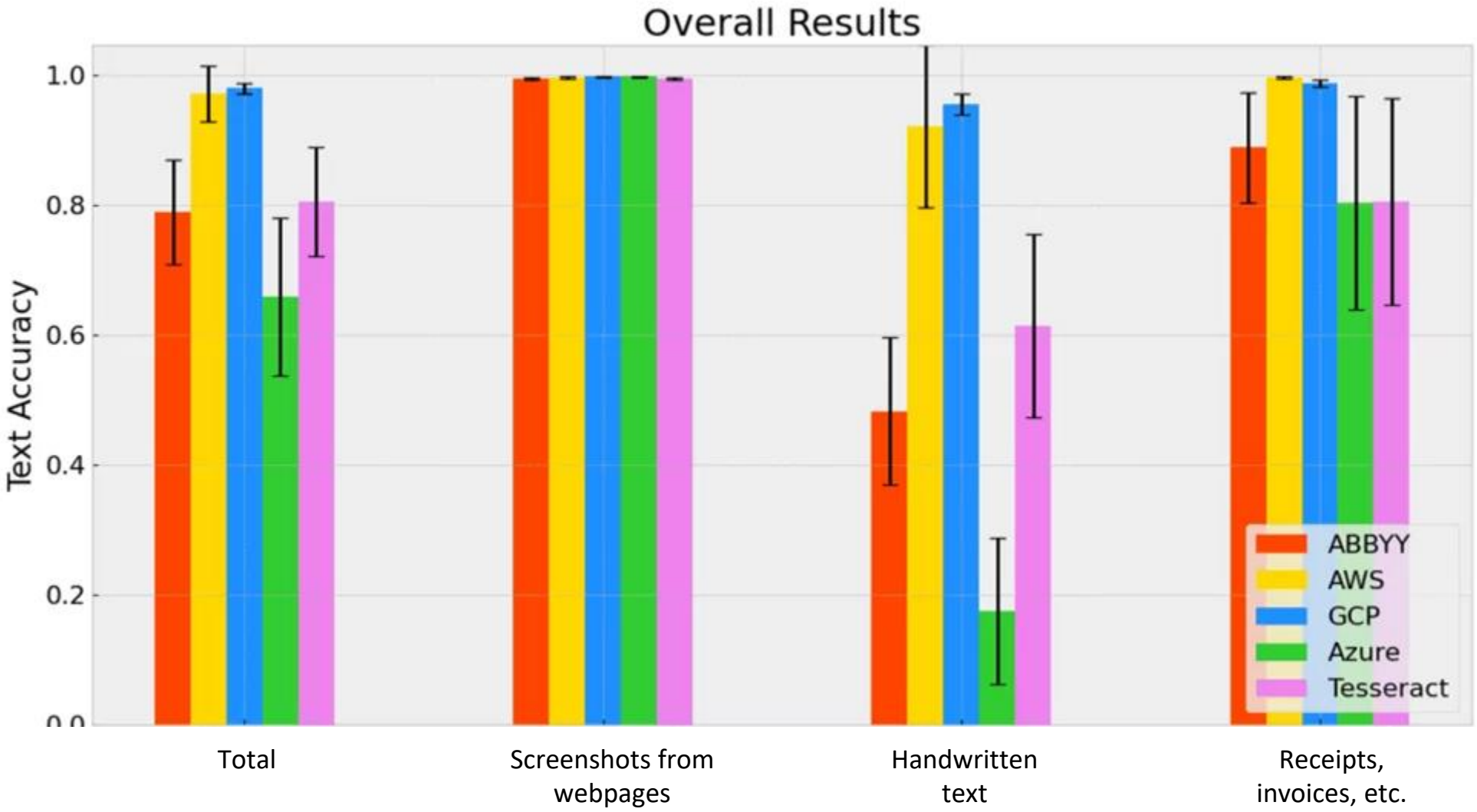
Amazon Textract

Google Cloud
Cloud Vision API

Microsoft Azure

Tesseract OCR

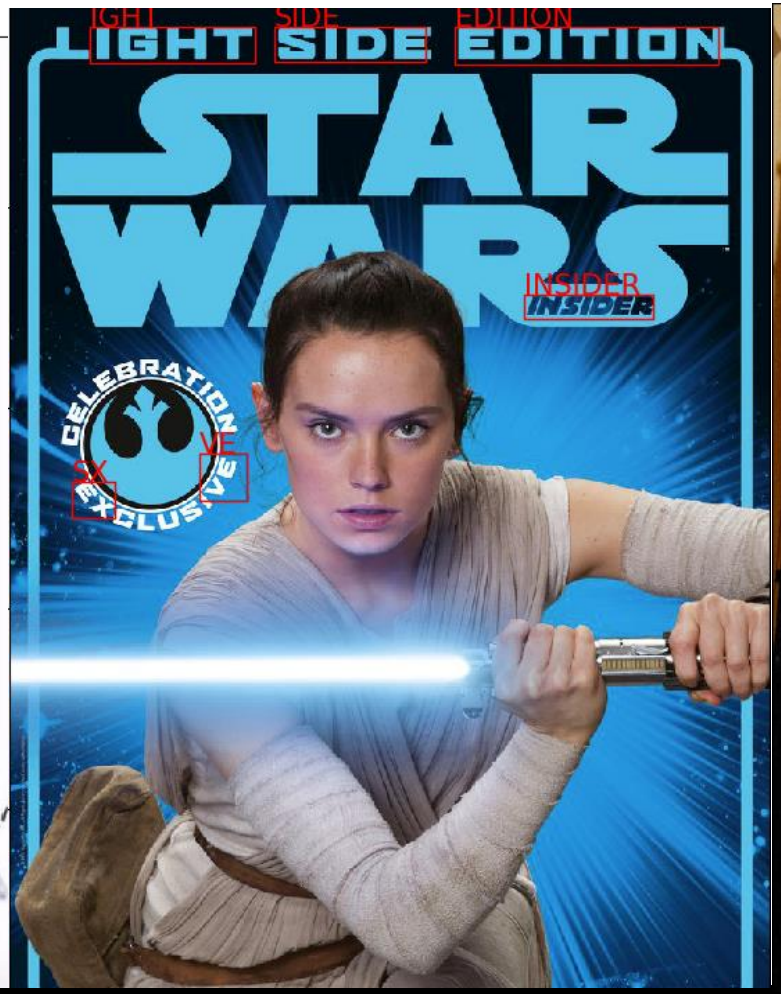
Commercial OCRs



source: https://research.aimultiple.com/ocr-accuracy/

Commercial OCRs

Some failure cases using Microsoft's OCR on scene-text data.



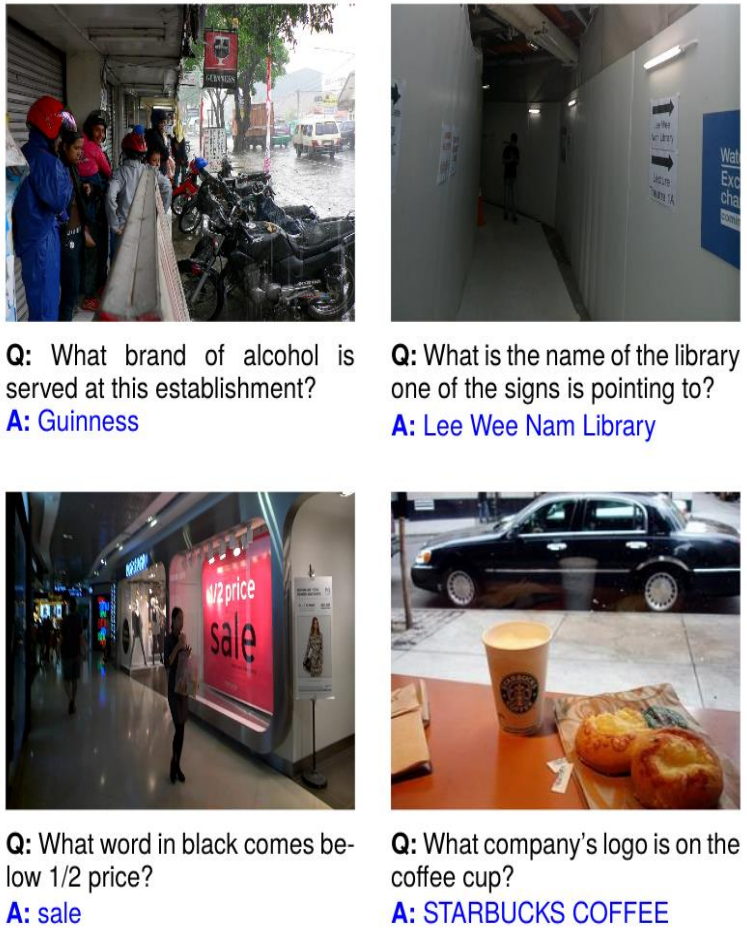
Reasoning with Text: Scene Text Visual Question Answering (ST-VQA)

VQA with image text

Visual Question Answering

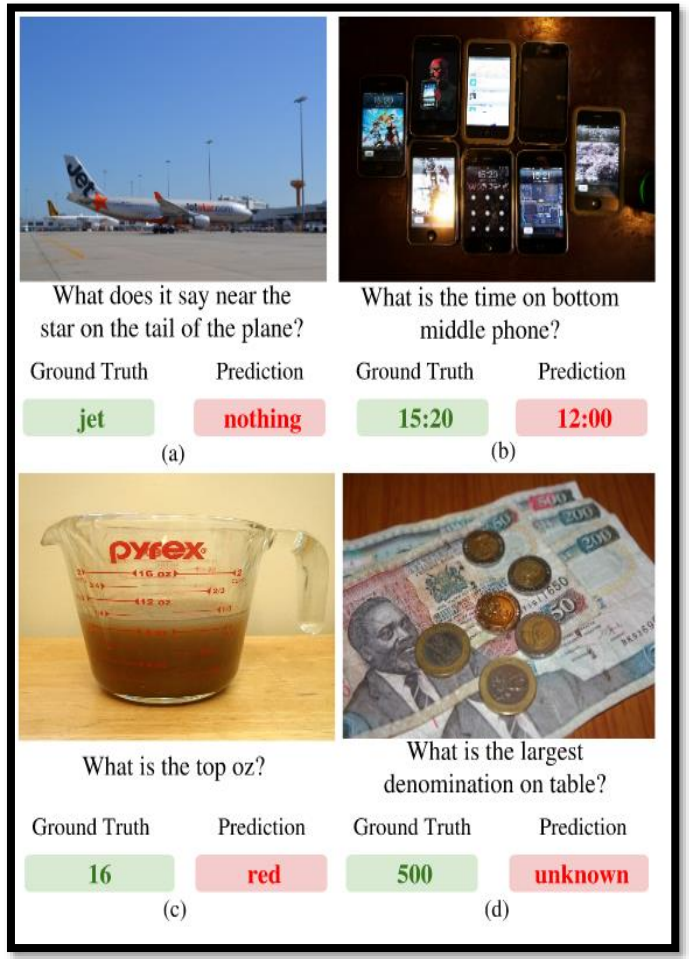
(Agarwal et al., 2015)

Scene Text VQA



(Biten et al., 2019)

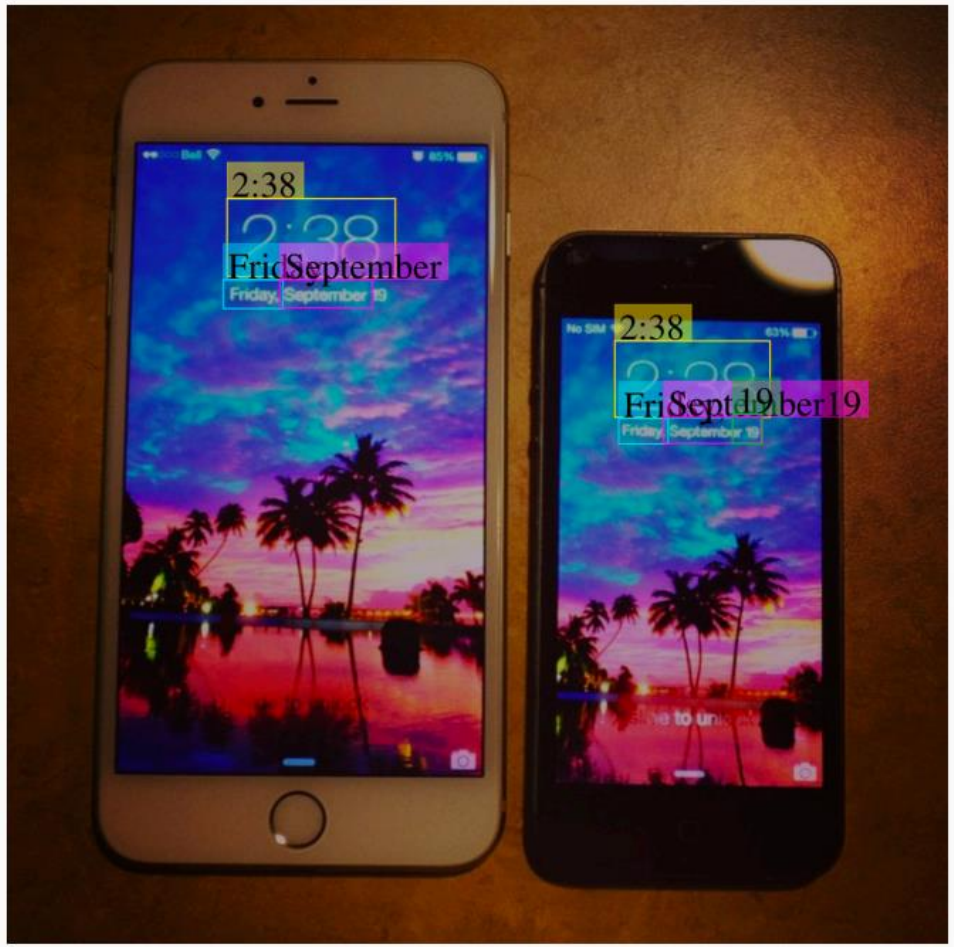
TextVQA



(Singh et al., 2019)

Challenges

- Open-Ended Vocabulary



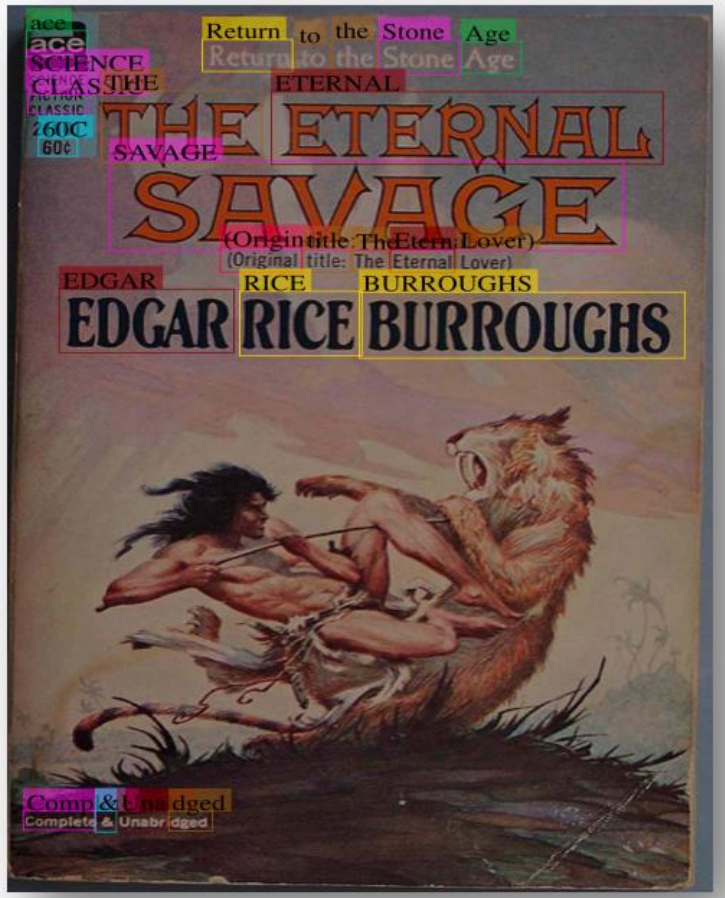
What time is it?

(2:38)

Slides taken from: https://drive.google.com/file/d/1ltTIyc6eubE7NerVPQI7sN7ECcoLhWkD/view?usp=sharing

Challenges

- Open-Ended Vocabulary
- Multi-word predictions



Who wrote this book?

(edgar rice burroughs)

Slides taken from: https://drive.google.com/file/d/1ltTIyc6eubE7NerVPQI7sN7ECcoLhWkD/view?usp=sharing

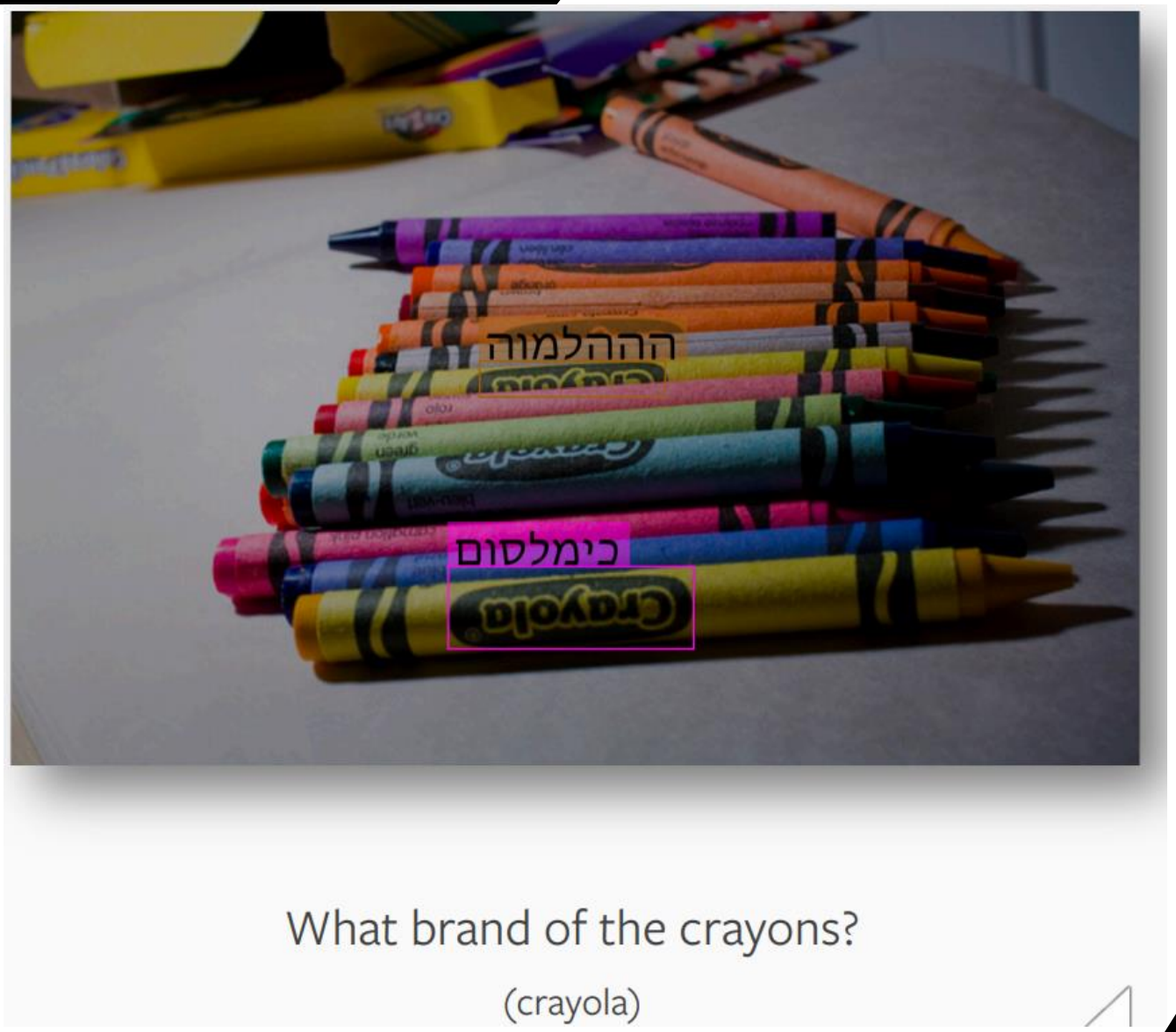
Challenges

- Open-Ended Vocabulary
- Multi-word predictions
- Text Detection

Slides taken from: https://drive.google.com/file/d/1ltTIyc6eubE7NerVPQI7sN7ECcoLhWkD/view?usp=sharing

Challenges

- Open-Ended Vocabulary
- Multi-word predictions
- Text Detection
- Text Extraction



Slides taken from: https://drive.google.com/file/d/1ltTIyc6eubE7NerVPQI7sN7ECcoLhWkD/view?usp=sharing

Challenges

- Open-Ended Vocabulary
- Multi-word predictions
- Text Detection
- Text Extraction
- Reasoning

Slides taken from: https://drive.google.com/file/d/1ltTIyc6eubE7NerVPQI7sN7ECcoLhWkD/view?usp=sharing

Challenges

- Open-Ended Vocabulary
- Multi-word predictions
- Text Detection
- Text Extraction
- Reasoning
- World Knowledge

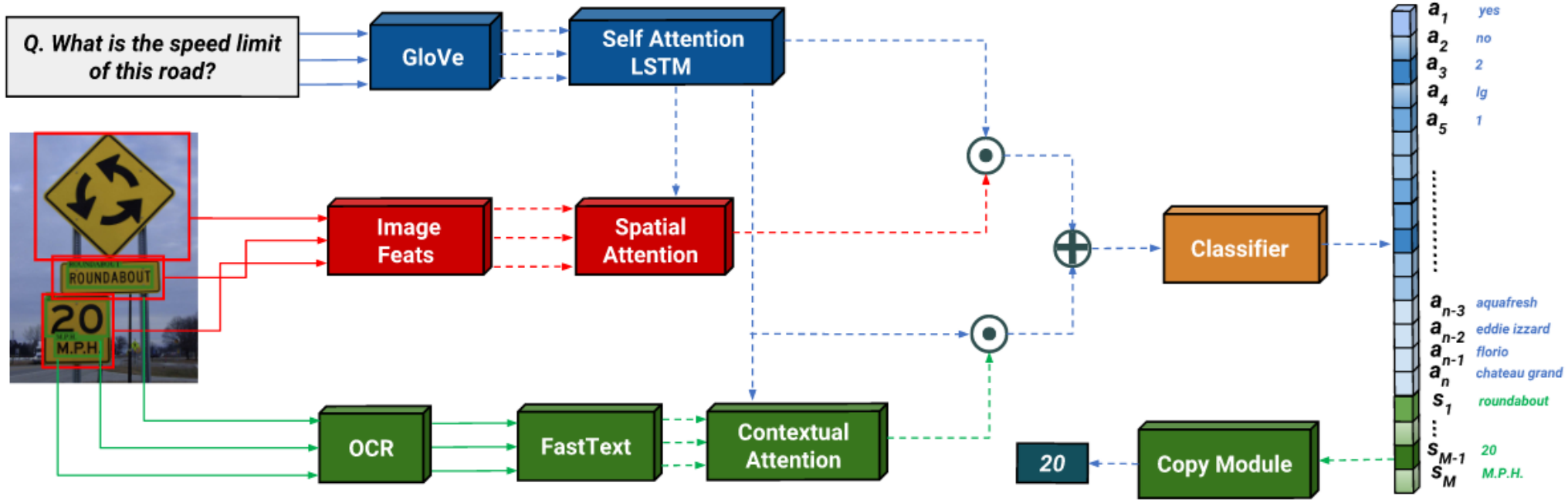
Which processor brand is featured on top left?

(intel)

Slides taken from: https://drive.google.com/file/d/1ltTIyc6eubE7NerVPQI7sN7ECcoLhWkD/view?usp=sharing

Look, Read, Reason & Answer (LoRRA)

- Three different branches for question, image and OCR with attention conditioned on the question. Feature fusion to get the answer
- Posed as a classification problem, using a pre-defined vocabulary words and OCR tokens. Only one word per answer
- OCR is integrated through the Copy Module, that selects one OCR word as the predicted answer. The classifier can predict the index of an OCR token instead of a vocabulary word and that word is copied as the answer,



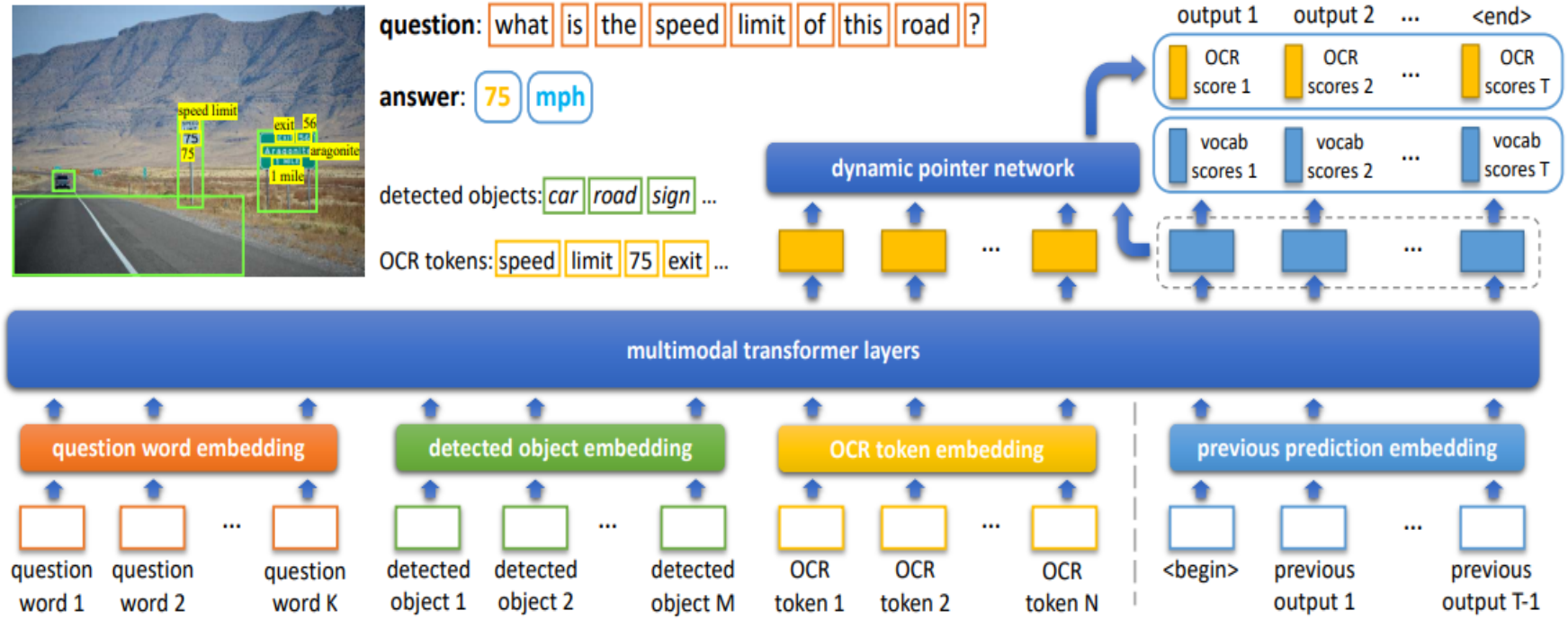
Test Accuracy ST-VQA: -

Test Accuracy TextVQA: 27.63%

Singh et al. "Towards vqa models that can read." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

M4C: multimodal scene text VQA

- A transformar encoder that jointly encodes question words, object features and OCR tokens
- Answer obtained through a autoregressive generative process. A transformar decoder generates one word at each time step taken into account the otuput of the encoder and previous generated words. Answer can be multi-word.
- OCR is integrated through the Dynamic Pointer Network that generates a score for each of the encoded OCR tokens. At each time step the decoder can select the next word among the pre-defined vocabulary words or the OCR tokens.



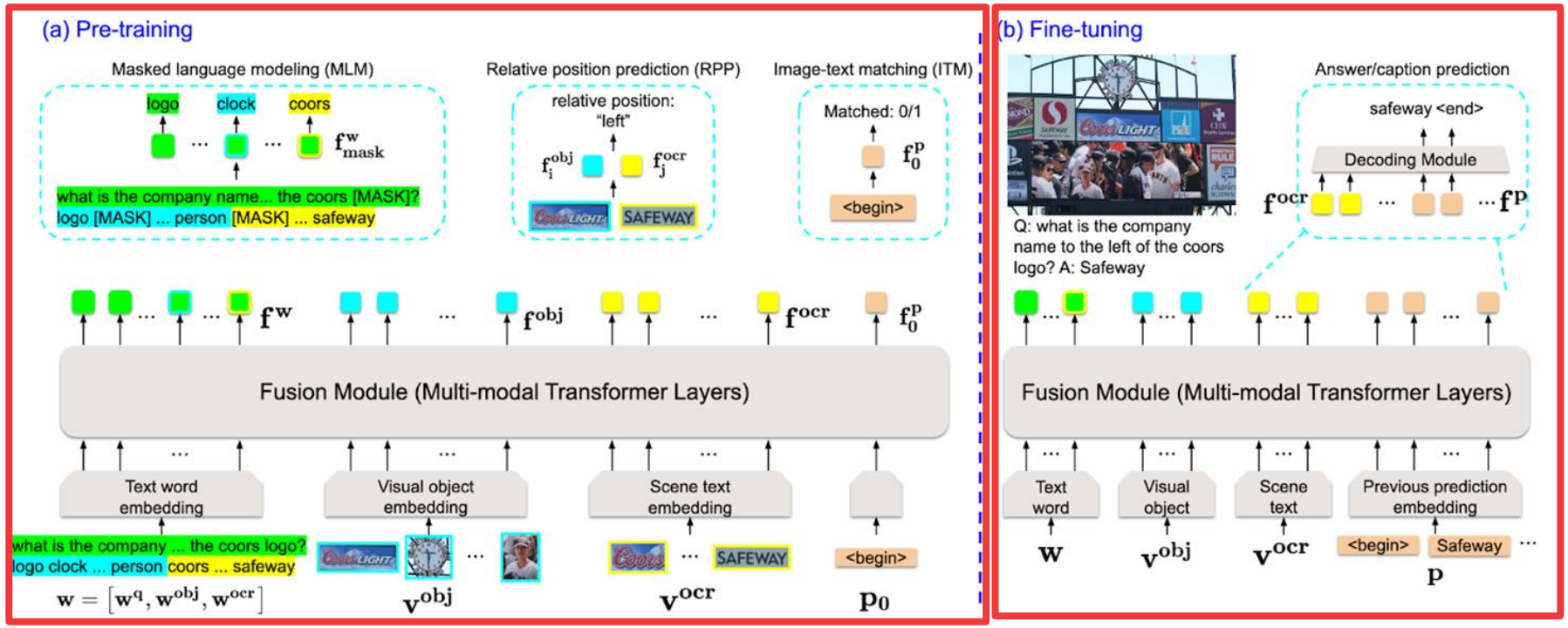
Test Accuracy ST-VQA: 38.05%

Test Accuracy TextVQA: 40.55%

Hu et al, "Iterative answer prediction with pointer-augmented multimodal transformers for TextVQA", CVPR 2020

Text Aware Pretraining (TAP)

- Same basic architecture as M4C with slight modifications: integrates all text tokens (question, object labels and OCR tokens) into a single text word channel and adds a specific channel with visual features of scene text
- Adds a pre-training stage to learn a better joint representation of question, image and OCR
  - Standard Masked Languagel Modelling on the joint text word channel (question, object labels and OCR)
  - Image Text Matching: predicting if randomly sampled text word and image channels are paired or not.
  - Relative Position Prediction: predicting relative spatial position of randomly sampled visual objects and OCR tokens



Test Accuracy ST-VQA: 50.83%

Test Accuracy TextVQA: 54.71%

Zhengyuan et al. "Tap: Text-aware pre-training for text-vqa and text-caption." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

LaTr

It realizes that most questions can be used just with text and spatial layout.
Visual information is necessary only a very few cases

Question: what is the website listed on the banner?

Global
Historias
HiperBarrio
Audiencia
Locales,
http://hiperbarrio.org/

Bag of Words

41.77%

HiperBarrio
Historias
Locales,
Audiencia
Global
http://hiperbarrio.org/

Ordered Bag of Words

2D Spatial Position
+
Words

2D Spatial Position
+
Words
+
Image

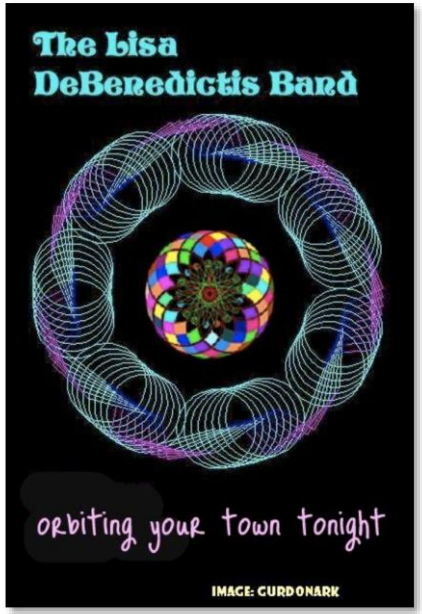
LaTr

It realizes that most questions can be used just with text and spatial layout.
Visual information is necessary only a very few cases

Question: what is the website listed on the banner?

The
Town
Band
Lisa
Orbiting
DeBenedictis
Your

The
Lisa
DeBenedictis
Band
Orbiting
Your
Town



Bag of Words

41.77%

Ordered Bag of Words

50.37%

2D Spatial Position
+
Words

2D Spatial Position
+
Words
+
Image

Biten et al, “LaTr: Layout-Aware Transformer for Scene-Text VQA”, CVPR 2022

LaTr

It realizes that most questions can be used just with text and spatial layout.
Visual information is necessary only a very few cases

Question: what is the name of the video game?

Push
Sega/
1966
play
Start
free

Sega
windows
button
OutRun
Sharp

OutRun
Push
Start
button
Sega
1966

Free
Play
Sharp
Sega/
windows

Bag of Words	Ordered Bag of Words	2D Spatial Position + Words	2D Spatial Position + Words + Image
41.77%	50.37%	51.22%	

Biten et al, “LaTr: Layout-Aware Transformer for Scene-Text VQA”, CVPR 2022

LaTr

It realizes that most questions can be used just with text and spatial layout.
Visual information is necessary only a very few cases

Question: what does it say on the shirt of the man in the white?

Qatar
LEP
bwin
Foundation
adidas

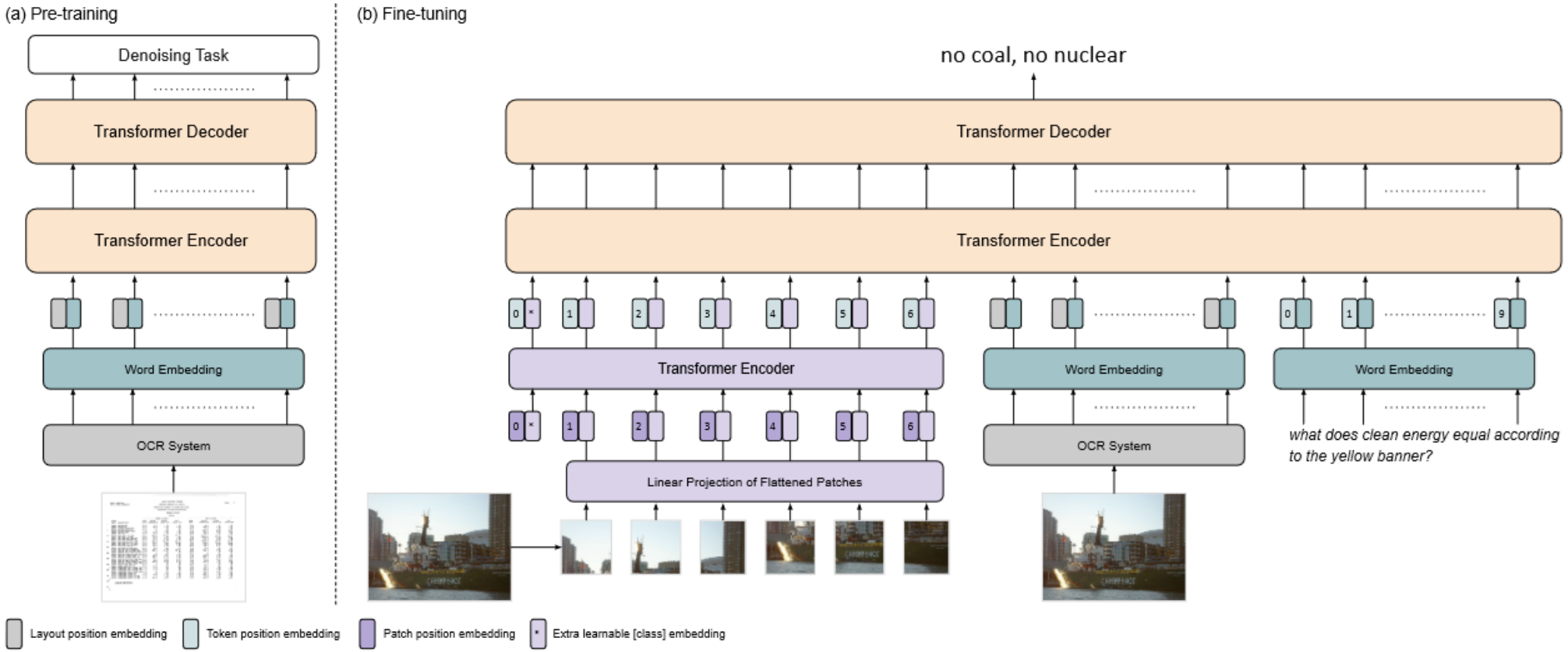
Foundation
Qatar
adidas
bwin
LEP

Bag of Words	Ordered Bag of Words	2D Spatial Position + Words	2D Spatial Position + Words + Image
41.77%	50.37%	51.22%	52.29%

Biten et al, “LaTr: Layout-Aware Transformer for Scene-Text VQA”, CVPR 2022

LaTr

- Architecture based on the T5 generative encoder/decoder model.
- They propose a new pre-training stage based on Standard MLM, but on document images where only text and spatial layout is considered

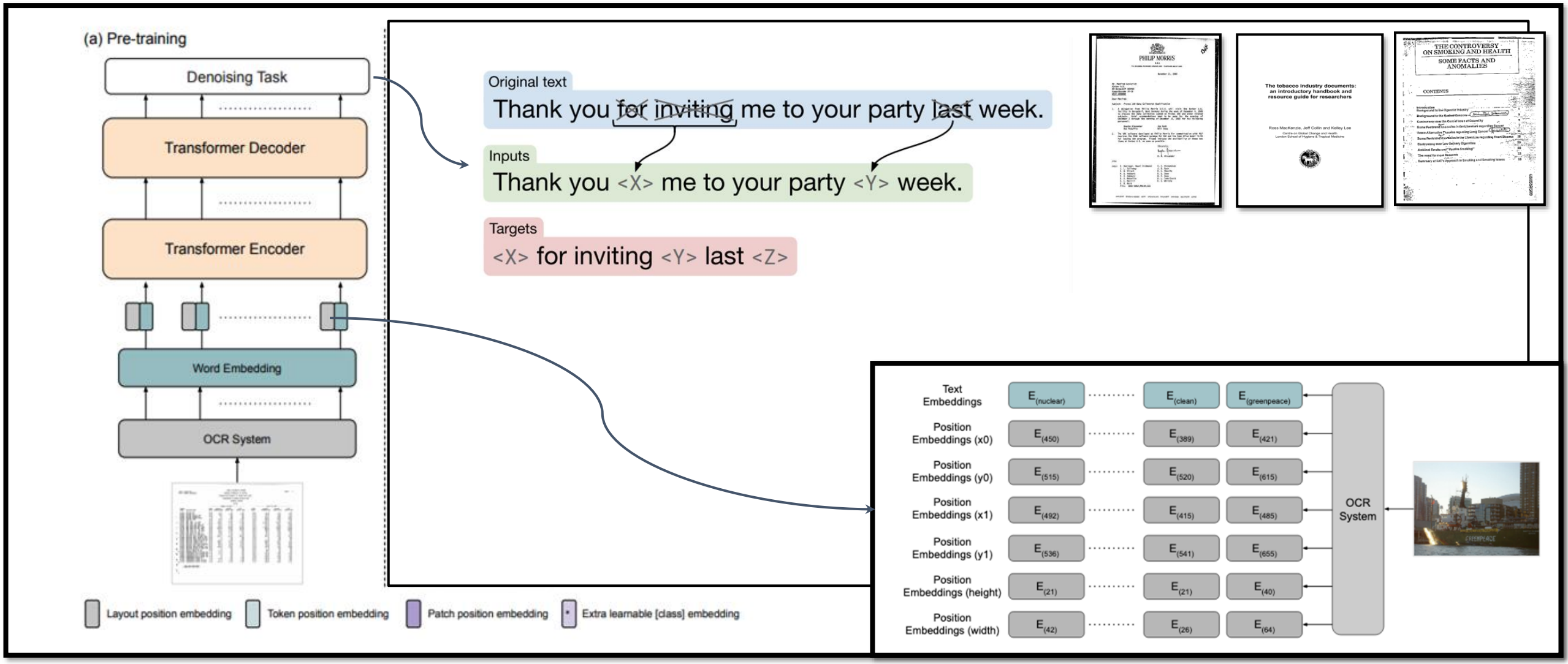


Test Accuracy ST-VQA: 61.64%

Test Accuracy TextVQA: 61.60%

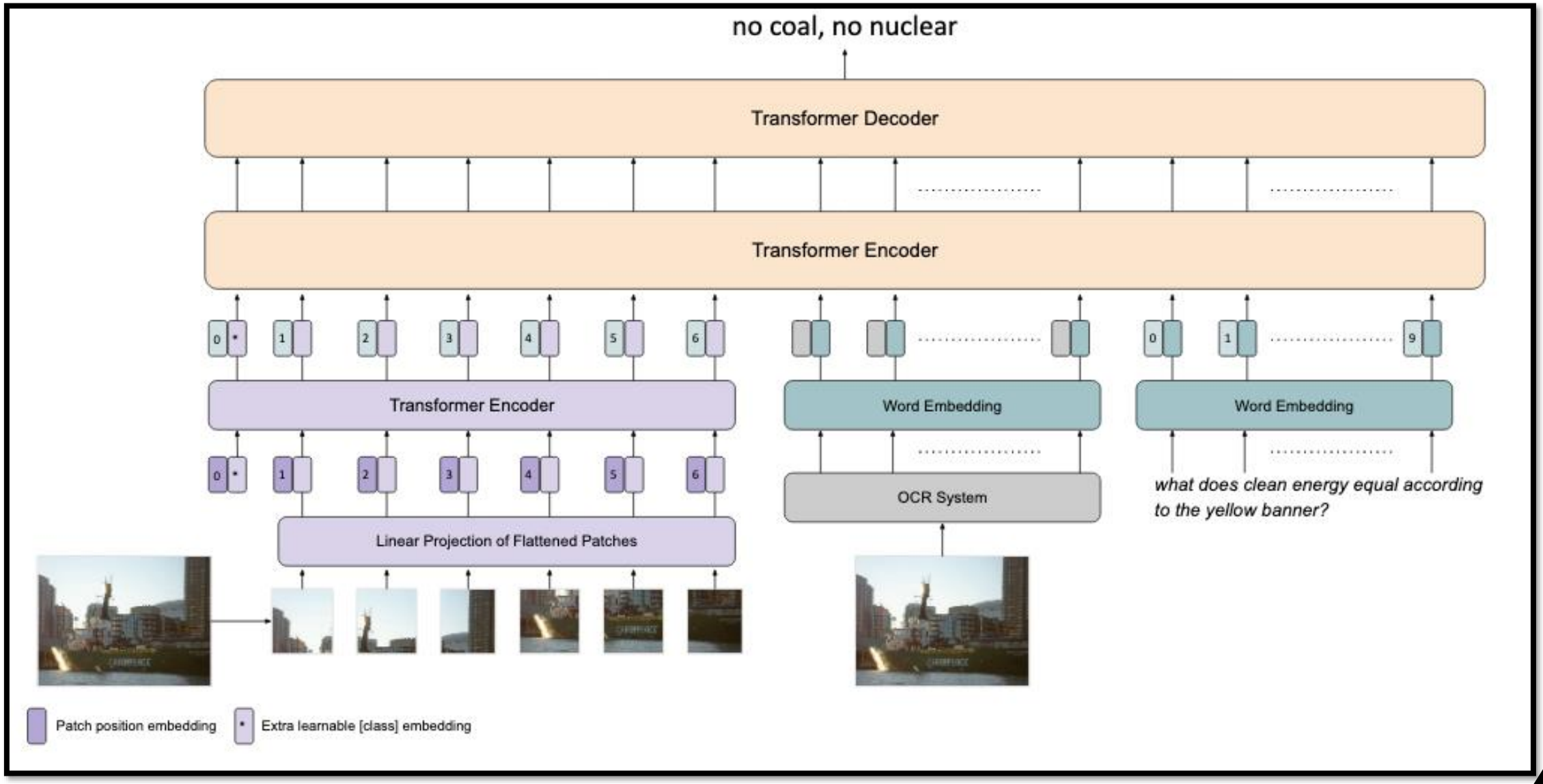
Biten et al, "LaTr: Layout-Aware Transformer for Scene-Text VQA", CVPR 2022

LaTr Pretraining



Biten et al, “LaTr: Layout-Aware Transformer for Scene-Text VQA”, CVPR 2022

Finetuning


no coal, no nuclear
Transformer Decoder
Transformer Encoder
Transformer Encoder
Word Embedding
Word Embedding
OCR System
Linear Projection of Flattened Patches
what does clean energy equal according to the yellow banner?
Patch position embedding
Extra learnable [class] embedding

Qualitative Results

What kind of food is on the menu?

M4C: tortas
LaTr: mexican
GT: mexican

What is the beer brand on the top shelf right side of the image?

M4C: choceto
LaTr: adams
GT: adams

What is this beverage called?

M4C: super lutica
LaTr: sambuca
GT: sambuca

What is the handwritten message?

M4C: you don't talk to...
LaTr: karl fogel
GT: karl fogel

What are the titles of these dvds?

M4C: the complete...
LaTr: the complete...
GT: south park