A Fine Matrix

Noah Rousell, Dom Chui, Anna Luiza Arantes

BROWN

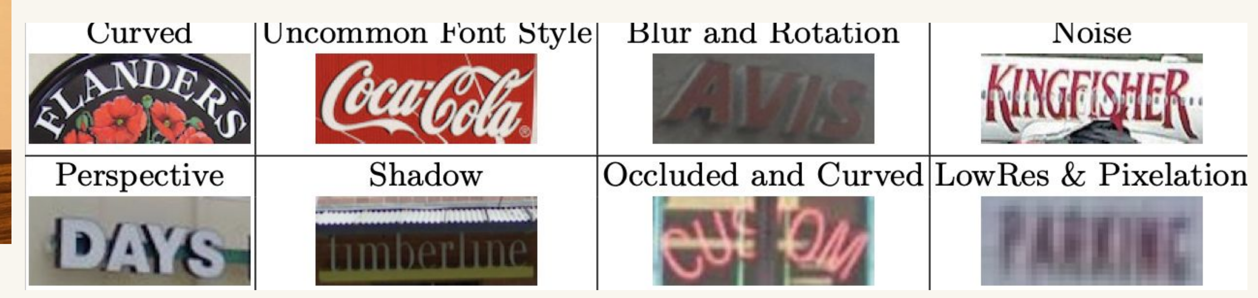
CSCI 1430

Goals and Motivations

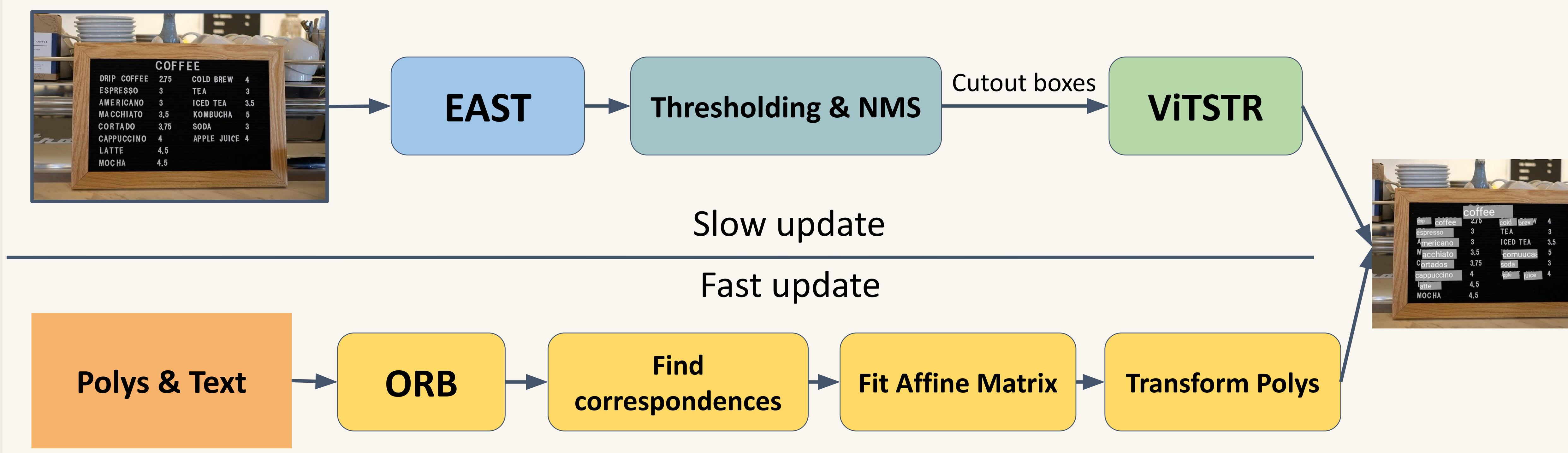
Our goal is to design a model that detects and translates text within different images, similar to a scoped-down version of the Google Lens Translate feature.

Importance: In an increasingly globalized world, there is a rising demand for the ability to comprehend and translate text across languages.

Challenge: in natural scenes, text lines may appear in a range of languages, styles, orientations, and other variations.



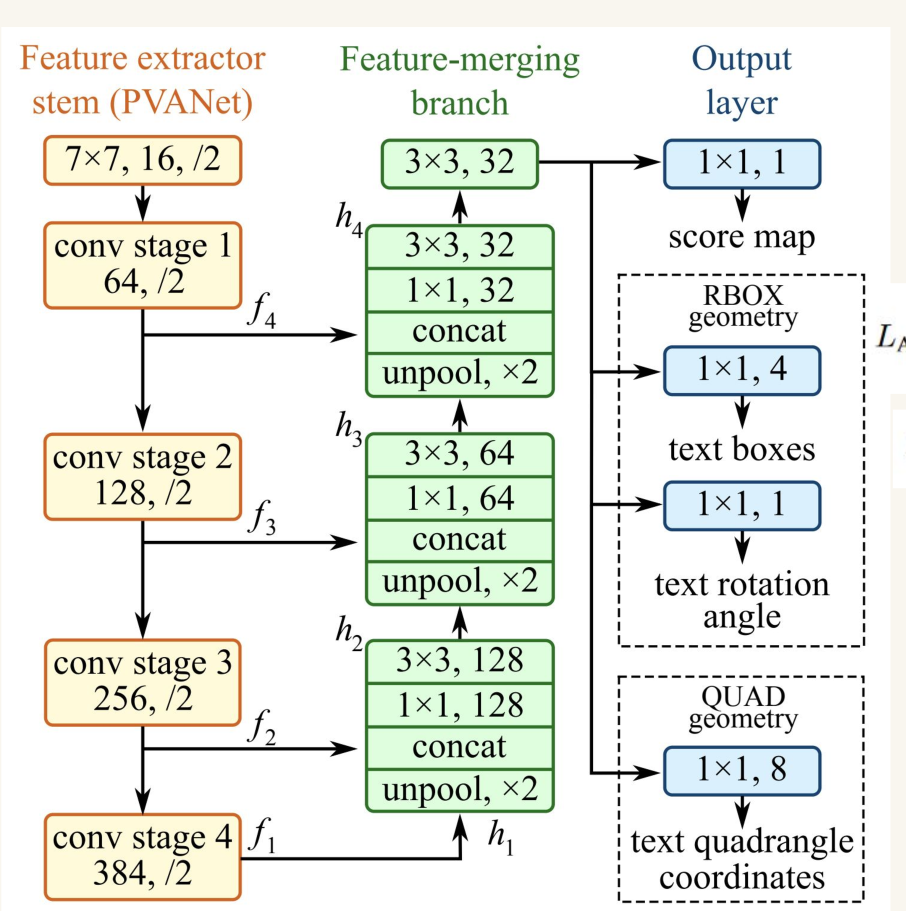
Architecture



EAST Detector

We implemented an EAST detector for text detection. The model makes use of a **fully convolutional network** (FCN) with **feature-merging at different scales** to produce text predictions (as rotated rectangles), which are then processed using thresholding and **Non-Maximum Suppression** to cull the poor candidates

Model Architecture:



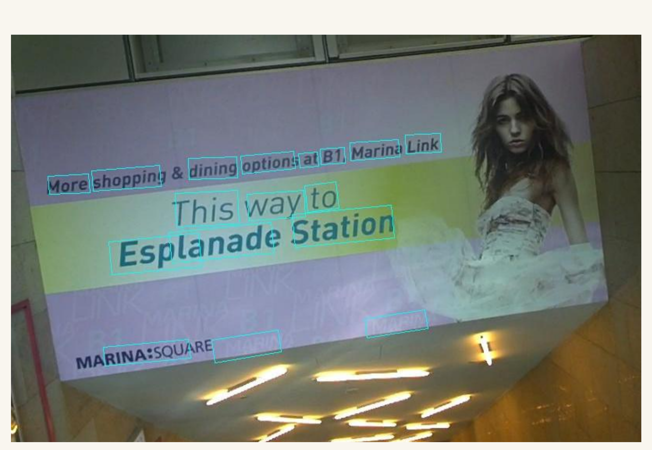
$$L = L_s + \lambda_g L_g$$

$$Dice = \frac{2\,|A \cap B|}{|A| + |B|}$$

$$L_{\mathrm{AABB}} = -\log \mathrm{IoU}(\hat{\mathbf{R}}, \mathbf{R}^*) = -\log \frac{|\hat{\mathbf{R}} \cap \mathbf{R}^*|}{|\hat{\mathbf{R}} \cup \mathbf{R}^*|}$$

$$L_\theta(\hat{\theta}, \theta^*) = 1 - \cos(\hat{\theta} - \theta^*).$$

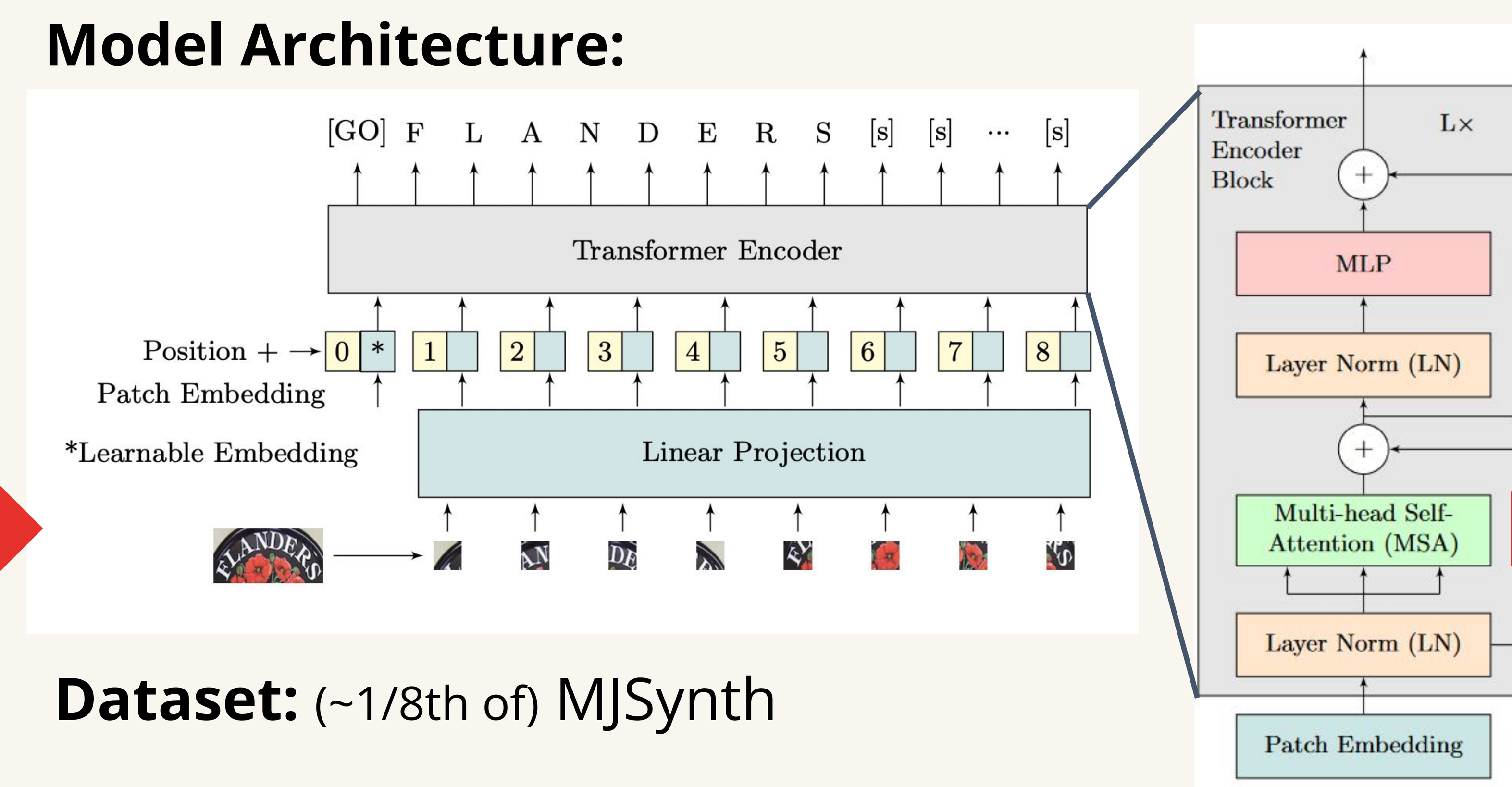
Dataset: ICDAR 2015



ViTSTR

For text recognition, we made use of a ViTSTR model, which uses the model weights of **Data efficient image Transformer**, which was trained used **hard-label distillation** with a strong **convnet teacher.**

Model Architecture:

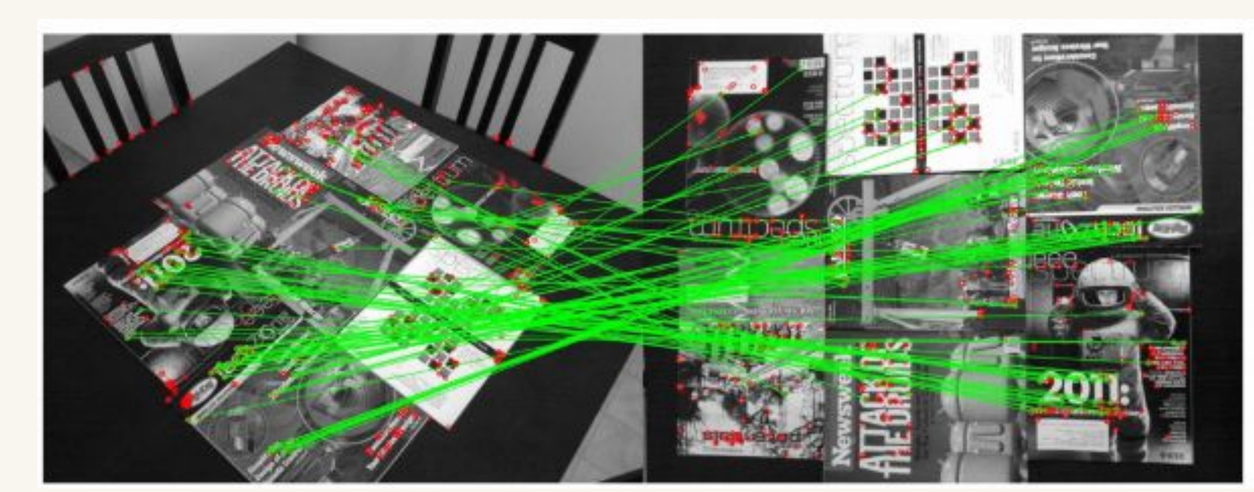


Dataset: (~1/8th of) MJSynth

Rendering & ORB

Once we have the polygons and text, we use the **Google Translate API** to receive translations in the target language. Then we render boxes and the translation on the source image, matching the scale and orientation.

As the goal is to have a live video, we need a faster way to update the polygons than a full forward pass of the neural network. We generate **ORB feature descriptors** (ORB utilizes the FAST algorithm to find a range of keypoints and BRIEF descriptors modified to account for rotation) and **find correspondences** to **fit an affine matrix** that describes the transformation between the slow update frame (t=0) and later frames (t=1,2,3,...). We apply this affine matrix to the polygons from t=0 to determine their likely location at the current time frame.



References

R. Atienza, "Vision Transformer for Fast and Efficient Scene Text Recognition." arXiv, May 18, 2021. doi: 10.48550/arXiv.2105.08582.
H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention." arXiv, Jan. 15, 2021. doi: 10.48550/arXiv.2012.12877.
E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *2011 International Conference on Computer Vision*, Nov. 2011, pp. 2564–2571. doi: 10.1109/ICCV.2011.6126544.
X. Zhou *et al.*, "EAST: An Efficient and Accurate Scene Text Detector." arXiv, Jul. 10, 2017. Accessed: Apr. 22, 2024. [Online]. Available: http://arxiv.org/abs/1704.03155

Acknowledgements

We would like to thank Professor Sridhar for his wonderful teaching and Brown CCV (Oscar) for compute resources!