Query-Centric Trajectory Prediction

https://www.youtube.com/watch?v=i46Sj0PUwyI

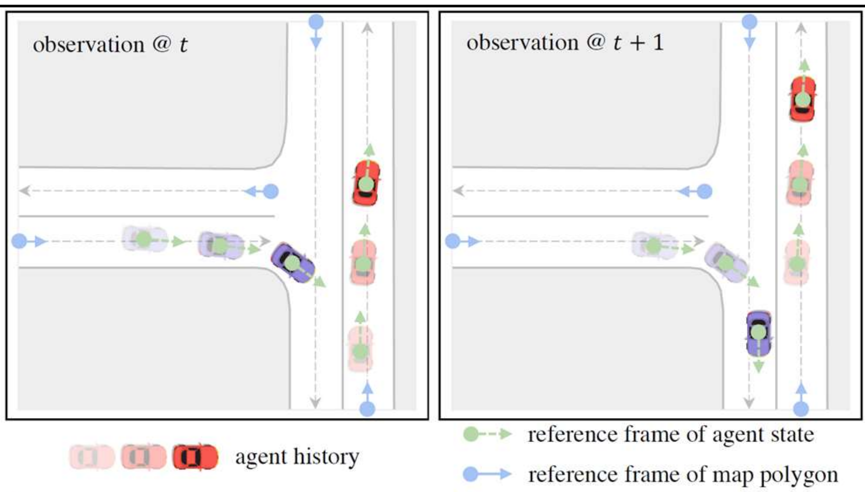


Figure 1. Illustration of our **query-centric reference frame**, where we build a local coordinate system for *each* spatial-temporal element, including map polygons and agent states at all time steps. In the attention-based encoder, all scene elements' queries are derived and updated in their local reference frames.

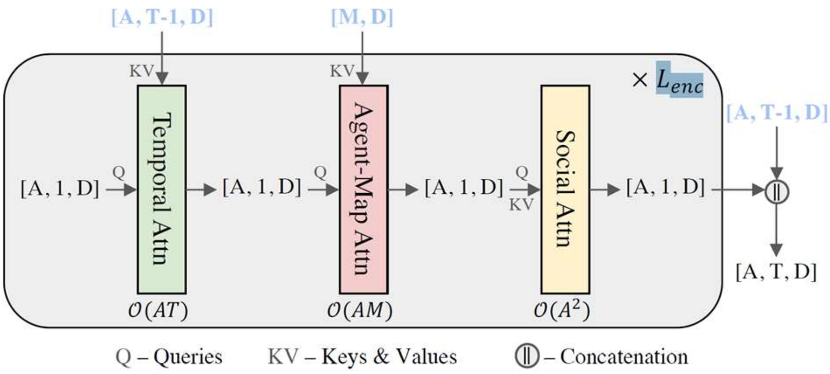


Figure 2. Overview of the **encoder** in an online mode. After reusing the encodings computed in previous observation windows (blue), the complexity of factorized attention goes from $\mathcal{O}(AT^2)+\mathcal{O}(ATM)+\mathcal{O}(A^2T)$ to $\mathcal{O}(AT)+\mathcal{O}(AM)+\mathcal{O}(A^2)$.

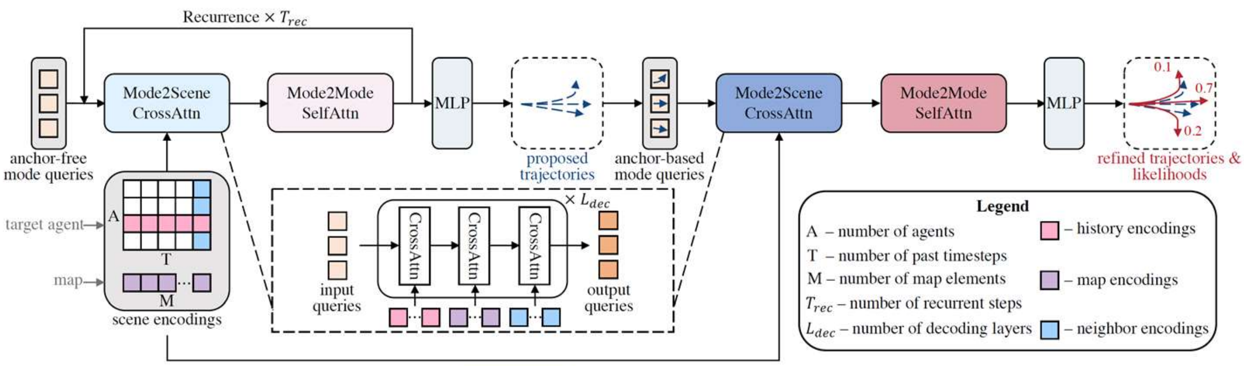
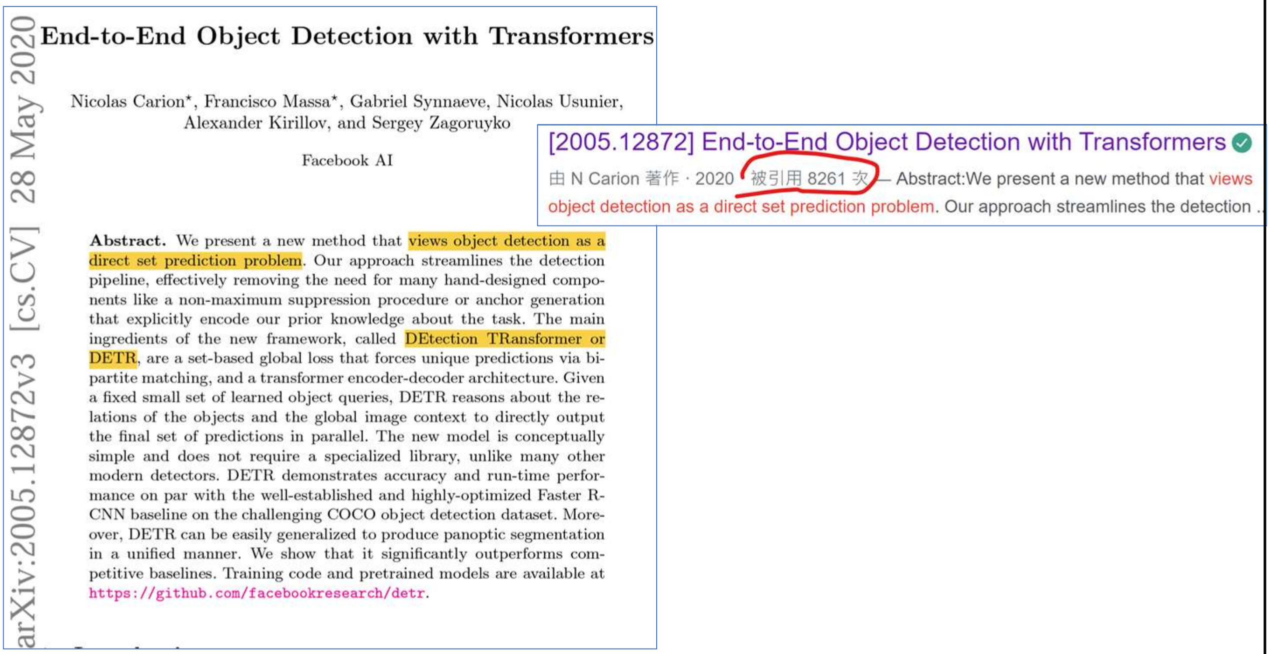


Figure 3. Overview of the **decoding pipeline**. An anchor-free module generates trajectory proposals *recurrently* based on the encoded scene context. These proposals act as the anchors in the refinement module, where an anchor-based decoder refines the anchor trajectories and assigns a probability score for each hypothesis.



arXiv:2005.12872v3 [cs.CV] 28 May 2020

End-to-End Object Detection with Transformers

Nicolas Carion*, Francisco Massa*, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko

Facebook AI

Abstract. We present a new method that views object detection as a direct set prediction problem. Our approach streamlines the detection pipeline, effectively removing the need for many hand-designed components like a non-maximum suppression procedure or anchor generation that explicitly encode our prior knowledge about the task. The main ingredients of the new framework, called DEtection TRansformer or DETR, are a set-based global loss that forces unique predictions via bipartite matching, and a transformer encoder-decoder architecture. Given a fixed small set of learned object queries, DETR reasons about the relations of the objects and the global image context to directly output the final set of predictions in parallel. The new model is conceptually simple and does not require a specialized library, unlike many other modern detectors. DETR demonstrates accuracy and run-time performance on par with the well-established and highly-optimized Faster R-CNN baseline on the challenging COCO object detection dataset. Moreover, DETR can be easily generalized to produce panoptic segmentation in a unified manner. We show that it significantly outperforms competitive baselines. Training code and pretrained models are available at https://github.com/facebookresearch/detr.

[2005.12872] End-to-End Object Detection with Transformers

由 N Carion 著作 · 2020 被引用 8261 次 — Abstract:We present a new method that views object detection as a direct set prediction problem. Our approach streamlines the detection ...

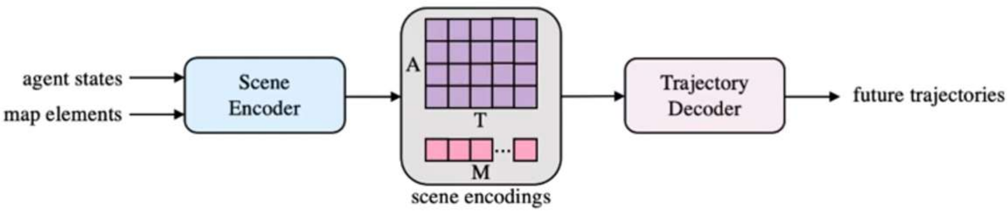
Agent-centric modeling

Re-normalizing and re-encoding the input

Research Gap

- Current methods need to redo calculations when observing changes, causing redundant computation in online prediction.
- Even with detailed scene information, current decoding methods have difficulty capturing multimodal outcomes in **long-term** predictions.

diverse



Method

- To achieve **faster** inference, we use a query-based approach for **encoding scenes** to reuse previous calculations. Sharing common scene features among target agents enables multi-agent trajectory decoding in parallel.

suggestions

- We begin by using anchor-free queries to create trajectory proposals in a recurrent fashion. This helps the model adapt to varying scene contexts for different prediction horizons. Next, a **refinement module** uses the trajectory proposals as starting points and employs anchor-based queries to fine-tune the trajectories.

step-by-step

Conclusions

- Providing adaptable, high-quality anchors to the **refinement module** enhances our query-based decoder's ability to handle diverse trajectory prediction outcomes.
- Achieves 1st on Argoverse 1 and Argoverse 2 motion forecasting benchmarks, outperforming all other methods by a large margin in all key metrics.
- Achieves real-time scene encoding and simultaneously decodes trajectories for multiple agents due to its **query-centric** design approach.

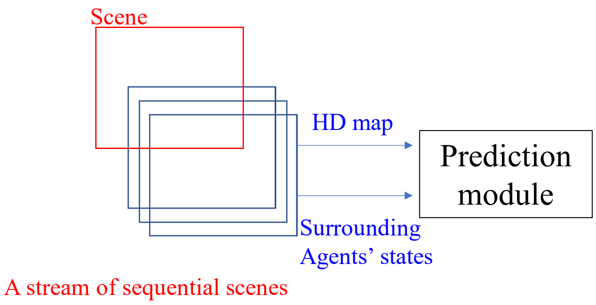
Limitation (I)

- Current approaches fail to process the heterogeneous traffic scenes efficiently.

Factorized attention-based Transformers

SOTA approaches 無法滿足real time需求， 所以難以應用在AV

Every minimal delay may lead to a catastrophic accident



Limitation (II)

Anchors are reference points used as a starting point for generating multiple possible future paths for agents. These anchors, along with associated probabilities, help account for the uncertainty and variability in predicting the future behavior of these agents.

- Uncertainty in the output of prediction becomes serious in long-term prediction.
- To ensure all potential behaviors are considered, a model should learn the diverse distribution, not just the most common outcome. This is challenging as each training sample records only a single possibility.

Solution

- Anchor-based methods use predefined anchors to achieve multimodal prediction, but the anchor quality strongly affects prediction accuracy.
- Anchor-free methods generate multiple hypotheses without constraints, which can lead to mode collapse and training instability.
- Combines both anchor-based and anchor-free approaches in the decoding process.
  - The anchor-free module creates adaptable anchors based on data, while the anchor-based module refines them according to the scene context.