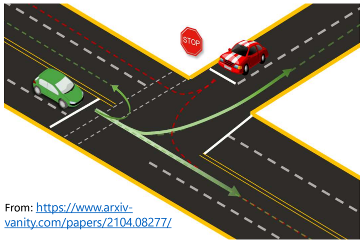
Can You Text What Is Happening? Integrating Pre-trained Language Encoders Into Trajectory Prediction Models for Autonomous Driving

Ali Keysan, Andreas Look, Eitan Kosman, Gonca Gürsun, Jörg Wagner, Yu Yao, Barbara Rakitsch

Introduction

Autonomous driving relies on predicting traffic behavior for safe navigation. While Large Language Models (LLMs), have shown promise in many different application areas, they haven't been explored for trajectory prediction.
Our work pioneers the use of text descriptions and pre-trained language encoders for trajectory prediction.

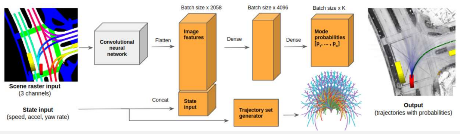


From: https://www.arxiv-vanity.com/papers/2104.08277/

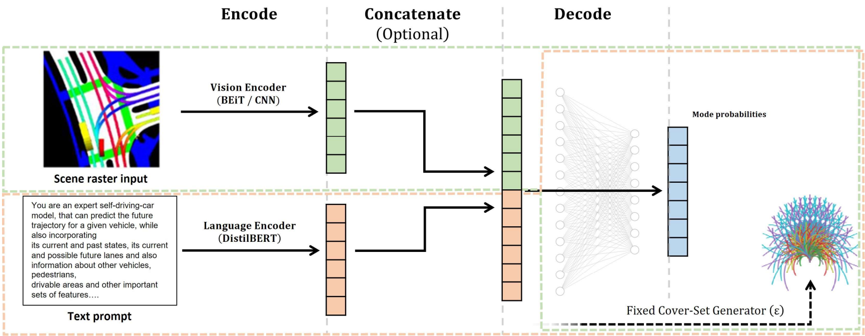
We find that text descriptions (i) offer a viable alternative to images and (ii) combining image and text encoders enhances performance.
We hope that our study encourages more research towards ultimately leading to more interpretable and expressive predictions for autonomous driving.

Background

Our approach builds on CoverNet, a standard trajectory prediction model. It employs an encoder-decoder architecture, initially processing scene representations as rasterized images. It then concatenates these embeddings with an agent's state vector and utilizes dense layers to generate predictions.
The trajectory prediction is cast as a classification task, selecting trajectories from a candidate set, $\mathcal{K}$, that effectively covers a broad range of possible trajectories while keeping it manageable in size.



Model Architecture



You are an expert self-driving-car model, that can predict the future trajectory for a given vehicle, while also incorporating
its current and past states, its current and possible future lanes and also information about other vehicles, pedestrians,
drivable areas and other important sets of features….

Flow of our Model. We encode the image that represents the rasterized scene and the text prompt with pre-trained models dedicated for each modality. If both input sources are used, we afterwards concatenate their embeddings. The result is fed into a decoder whose final layer picks the target trajectory from a pre-generated trajectory set.

Future Work

Our work sets the stage for future research in autonomous driving trajectory prediction.

- We foresee performance enhancements through alternative decoders that allow for expressive output representations.
- We anticipate that performance can be further increased by moving to larger models with over 10 billion parameters, coupled with parameter-efficient fine-tuning and soft prompting.
- Adopting joint image-and-text encoders promise substantial gains by capturing inter-modal relationships effectively.
- Our language models could generate auxiliary textual output for enhanced interpretability and facilitate scenario-specific instructions in traffic simulations.

Experiments

Method	Image	Text	#Modes	minADE_1	minADE_5	minADE_{10}	MissRate_1	MissRate_5	MissRate_{10}	minFDE_1
Constant velocity & yaw				4.61	4.61	4.61	0.91	0.91	0.91	11.21
Physics oracle				**3.70**	3.70	3.70	**0.88**	0.88	**0.88**	**9.09**
CoverNet (32M)	✓		64	5.16	2.41	2.18	N/A	0.92	N/A	10.84
CoverNet (34M)	✓		415	5.07	**2.31**	**1.76**	N/A	0.83	N/A	10.62
CoverNet (41M)	✓		2206	5.41	2.62	1.92	N/A	**0.76**	N/A	11.36
ResNet-50 (24M)	✓		64	5.07	2.56	2.20	0.95	0.92	0.92	10.61
ResNet-50 (24M)	✓		415	4.80	2.45	1.86	0.94	0.82	0.76	10.18
ResNet-50 (28M)	✓		2206	5.31	2.80	2.11	**0.93**	0.76	0.64	11.22
ResNet-152 (58M)	✓		64	4.86	2.47	2.17	0.95	0.92	0.92	10.15
ResNet-152 (59M)	✓		415	**4.51**	**2.33**	**1.80**	**0.93**	0.81	0.76	**9.57**
ResNet-152 (63M)	✓		2206	4.72	2.58	1.94	**0.93**	**0.75**	**0.63**	10.05
BEiT-B (86M)	✓		64	4.31	2.32	2.12	0.95	0.92	0.92	9.12
BEiT-B (86M)	✓		415	**3.92**	**1.98**	**1.57**	0.92	0.79	0.74	**8.46**
BEiT-B (88M)	✓		2206	4.20	2.29	1.75	**0.91**	**0.72**	**0.59**	9.22
$\text{DistilBERT}_{discr.}$ (67M)		✓	64	4.58	2.42	2.18	0.95	0.92	0.92	10.25
$\text{DistilBERT}_{discr.}$ (67M)		✓	415	4.31	2.24	1.74	0.92	0.80	0.75	9.97
$\text{DistilBERT}_{discr.}$ (69M)		✓	2206	4.86	2.80	2.11	**0.91**	**0.70**	0.57	11.30
DistilBERT (67M)		✓	64	4.45	2.39	2.16	0.95	0.92	0.92	9.94
DistilBERT (67M)		✓	415	**4.23**	**2.20**	**1.70**	0.93	0.80	0.75	**9.81**
DistilBERT (69M)		✓	2206	4.56	2.55	1.94	**0.91**	**0.70**	**0.56**	10.57
BEiT-B-DistilBERT (159M)	✓	✓	64	3.93	2.23	2.10	0.94	0.92	0.92	8.50
BEiT-B-DistilBERT (160M)	✓	✓	415	**3.62**	**1.87**	**1.49**	0.92	0.78	0.73	**8.09**
BEiT-B-DistilBERT (168M)	✓	✓	2206	3.73	2.00	1.53	**0.90**	**0.66**	**0.52**	8.41