Trajectory Prediction for Autonomous Driving

Milica Vukasinovic, Antoine Dávid

Context and Motivation

The aim of this project is to achieve a low trajectory prediction error — i. e. minimizing the Average Displacement Error (ADE), which remains a challenge especially due to complex urban environments.

- Autonomous driving systems require accurate predictions of future trajectories for safe and efficient planning
- Deep learning models have enabled significant progress in multi-task learning, where semantic segmentation, depth estimation, and trajectory prediction are jointly optimized

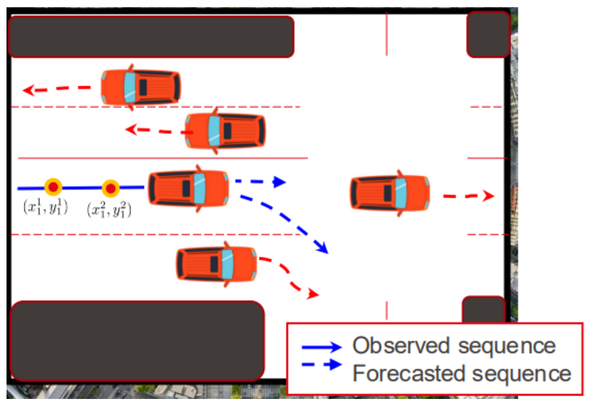

Figure 1. Applications of multi-task deep learning in autonomous driving [1].

This project consists of three milestones:

- **Milestone 1:** Implement a basic end-to-end planner that predicts ego-vehicle trajectories from RGB images, driving commands, and motion history.
- **Milestone 2:** Improve planning accuracy by incorporating multi-task learning with auxiliary perception tasks like semantic segmentation and depth estimation.
- **Milestone 3:** Enhance sim-to-real generalization by evaluating the planner in real-world settings without perception supervision and using data augmentation.

Data Overview

The given dataset consists of momentary sensor inputs (a front-facing RGB image, a depth image, a semantic segmentation map, vehicle trajectory). It is a curated subset of the nuPlan dataset [2].

- Each data sample includes past ego-vehicle poses and future ground-truth trajectories, enabling the supervised learning
- Semantic segmentation and depth maps to support multi-task learning across perception and planning
- The data is collected from diverse driving scenarios to improve generalization

Figure 2. Camera input sample with overlaid semantic and depth map [2].

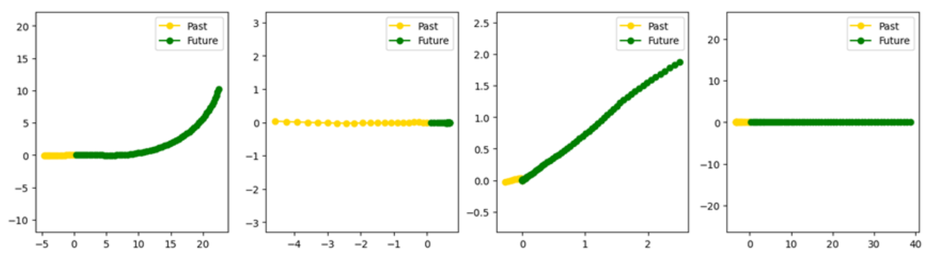

Figure 3. Ground-truth vehicle trajectory (in red) overlaid on scene [2].

Method Overview

Milestone 1:

- Uses an end-to-end model using an EfficientNet-B0 backbone pretrained on ImageNet to extract features from RGB camera images.
- The extracted visual features processed with the vehicle motion history through an MLP decoder.
- Here we used Smooth L1 loss, a LR of 1e-3, ADAM and a scheduler.

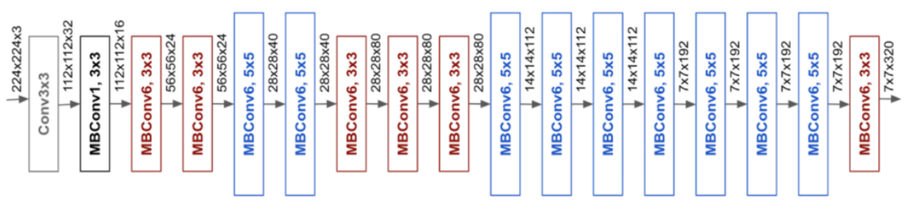

Figure 4. EfficientNet architecture [4].

Milestone 2:

- We use a ResNet34 as backbone instead of EfficientNet as we hit a ceiling in optimizing the Efficientnet.
- spatial attention block based on multi-head self-attention is added to the features enhancing spatial feature representation.
- We include auxiliary decoders for depth and semantic segmentation to train using auxiliary losses for depth and semantic.
- We again used ADAM, a scheduler and a LR = 2e-3

Method Overview

Milestone 3:

- We switch to a pretrained Vision Transformer (ViT-B/16) for improved spatial encoding.
- After 50 epochs we unfreeze the ViT backbone for fine-tuning.

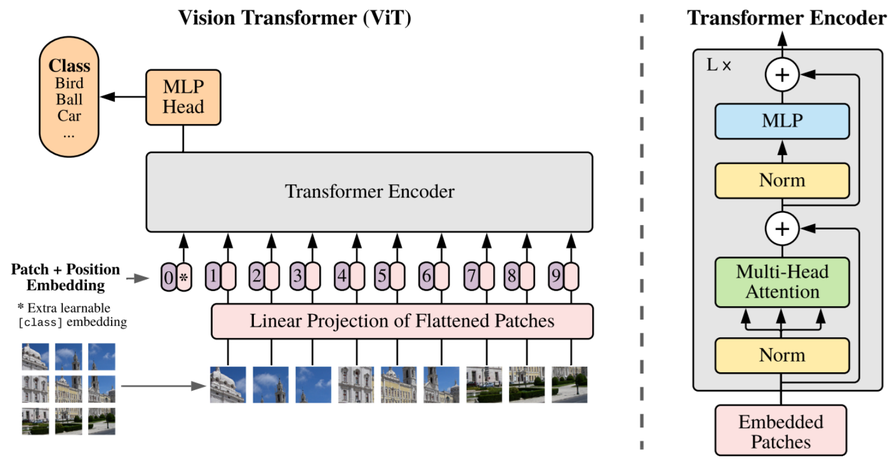

Figure 5. Visual Transformer architecture [3].

- We rely on a straight forward MLP for jointly decoding the features and motion history.
- We propose a combined loss function using a MSE of the positional predictions and a heading loss as $1 - \cos(\hat{\theta} - \theta)$ Thus utilizing the heading data for improved angular alignment.
- After each epoch we employ data augmentation including random horizontal flipping, rotation, cropping and color jittering of brightness, contrast, saturation and hue.
- In accordance with the pretrained ViT we preprocess the camera RGB input by resizing it to a shape of $[3, 224, 224]$ and normalizing it using the ImageNet mean and std.

Results

- Milestone 1 achieved an ADE of 1.82
- Milestone 2 achieved an ADE of 1.77
- Milestone 3 achieved an ADE of 1.72

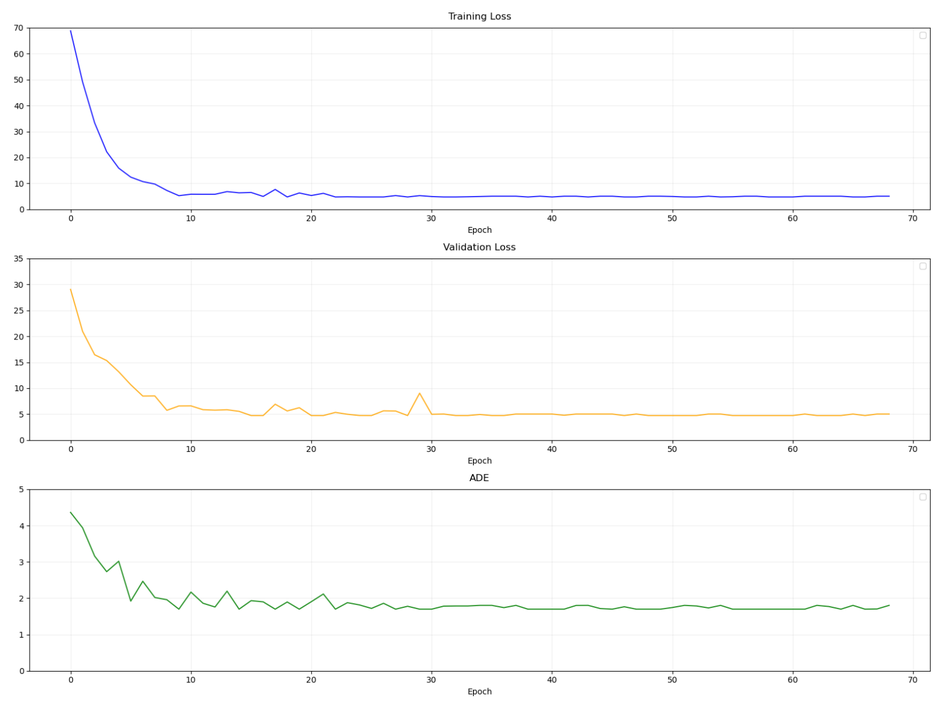

Figure 6. Training, Validation Loss and ADE

Discussion

- Our results in confirm that ViT is a viable approach for trajectory prediction in autonomous driving, achieving the 1.8 benchmark of the Average Displacement Error.
- However, the model is rather large and it thus takes a lot of compute for training it. Comparable results are reachable with smaller architectures.
- Further improvements can be made in our algorithms like:
  - improving the data augmentation (currently the trajectory is not flipped when the camera image is)
  - smoothing the predicted trajectory (current predictions are noisy and not drivable)
  - further analyzing the driving history feature, to incorporate extra features as speed and velocity as input
  - to further experiment with a wider range of learning rates, which was difficult due to the computing power required for training. Seperate learning rates for the ViT and the MLPs might be further beneficial for the ViT is pretrained.

References

Alexandre Alahi. Lecture 1: Introduction to deep learning for autonomous vehicles. Lecture, Deep Learning for Autonomous Vehicles (CIVIL-459), EPFL, February 2025. Available at `https://mediaspace.epfl.ch/playlist/dedicated/30913/0_zp56e5tk/`.

Holger Caesar, Vivek Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qing Xu, Anush Krishnan, Yung-Hsu Yang Pan, Giancarlo Baldan, and Oscar Beijbom. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. `https://www.nuscenes.org/nuplan`, 2021. Accessed: 2025-05-27.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2021. Published as a conference paper at ICLR 2021.

Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 6105–6114. PMLR, 2019. arXiv:1905.11946.