Chasing Clouds with Donkeycar: Holistic Exploration of Edge and Cloud Inferencing Trade-Offs in E2E Self-Driving Cars

Kyle Zheng, Kate Keahey (advisor), Alicia Esquivel Morel (advisor)
Modesto Junior College, The University of Chicago, Argonne National Laboratory, University of Missouri - Columbia

National Science Foundation · Modesto Junior College · The University of Chicago · Argonne National Laboratory · College of Engineering University of Missouri · Chameleon www.chameleoncloud.org

Cloud-Aided Real-time Inferencing Framework

- **Edge inference in autonomous vehicles**, while reliable, is constrained by resources
- **Cloud-assisted frameworks** that supplement edge devices **can introduce modular solutions to avoid potential bottlenecks** in vehicle actuation
- It is important to **analyze this possible solution** with regard to the problems of latency and resource trade-offs

Reproducing NVIDIA Paper [1] - Conceptual Reproduction

- **Conclusion of Paper is Demonstrated**
  - E2E Learning without decomposing the problem
  - Convolutional Neural Networks are able to abstract salient features (such as where the road is and avoid obstacles) from image input and use them to actuate a car without feature extraction being made into a separate step
- **Data Gathering**
  - Manually cleaned for undesirable behavior like driving off the road
- **Metric Used:** Autonomy Score is Analogous calculated with equation:

$$\text{autonomy} = \left(1 - \frac{\text{interventions} \times 6\,\text{seconds}}{\text{total time}}\right) \times 100\%$$

[1] Bojarski, Mariusz, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel et al. "End to end learning for self-driving cars." arXiv preprint arXiv:1604.07316 (2016).

Methodology

- **Hardware**
  - Scale car versus Real Car
  - RPi4 versus NVIDIA DRIVE
- **Architecture of the Neural Networks are changed**
  - Demonstrates that the conclusion of the NVIDIA paper is applicable to various architectures and neural network types
- **Amount of training data**
  - 72 hours for NVIDIA versus 1 hour for Reproduction
- **Frame Operation:**
  - NVIDIA captured images at 30 FPS, but the Reproduction uses 20 FPS

Motivation

- There is a **limited amount of resources** for every part of Donkeycar to use and high CPU utilization can bottleneck operations
- **The high, on-edge resource utilization** can cause under-performance of the various Donkeycar

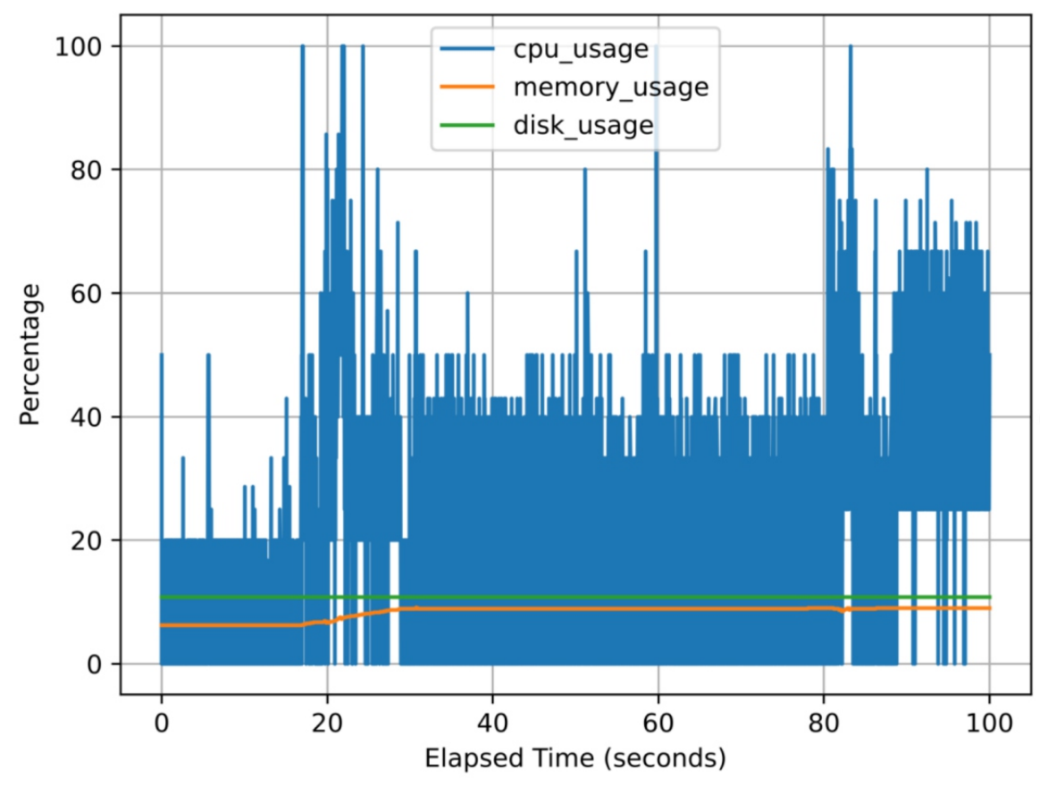

Fig. 1 Linear Model Resource Utilization

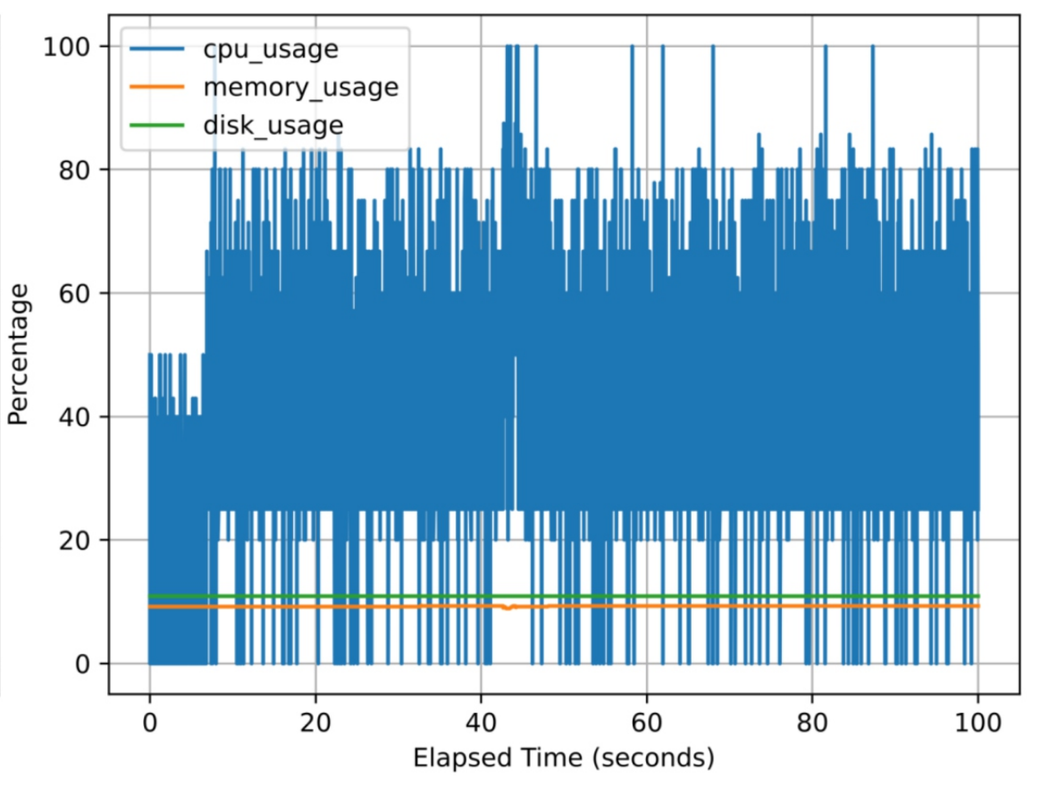

Fig. 2 LSTM3 Model Resource Utilization

- The resources on the **edge are also not able to operate** at the optimal vehicle loop frequence (20 loops)

- Even the fastest model on the **edge can only produce around 18 inferences on the RPi4**, whereas as the slowest model can produce 40 inferences on a RTX 6000 Inference

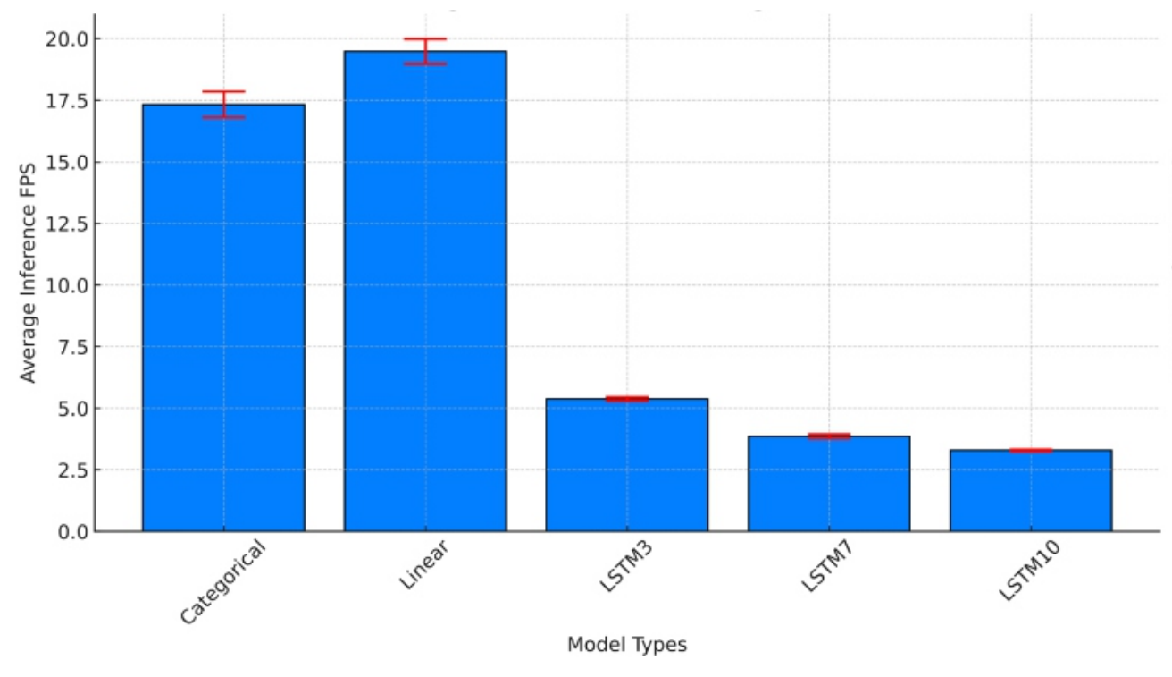

Fig. 3 Average Inference FPS for Edge Models

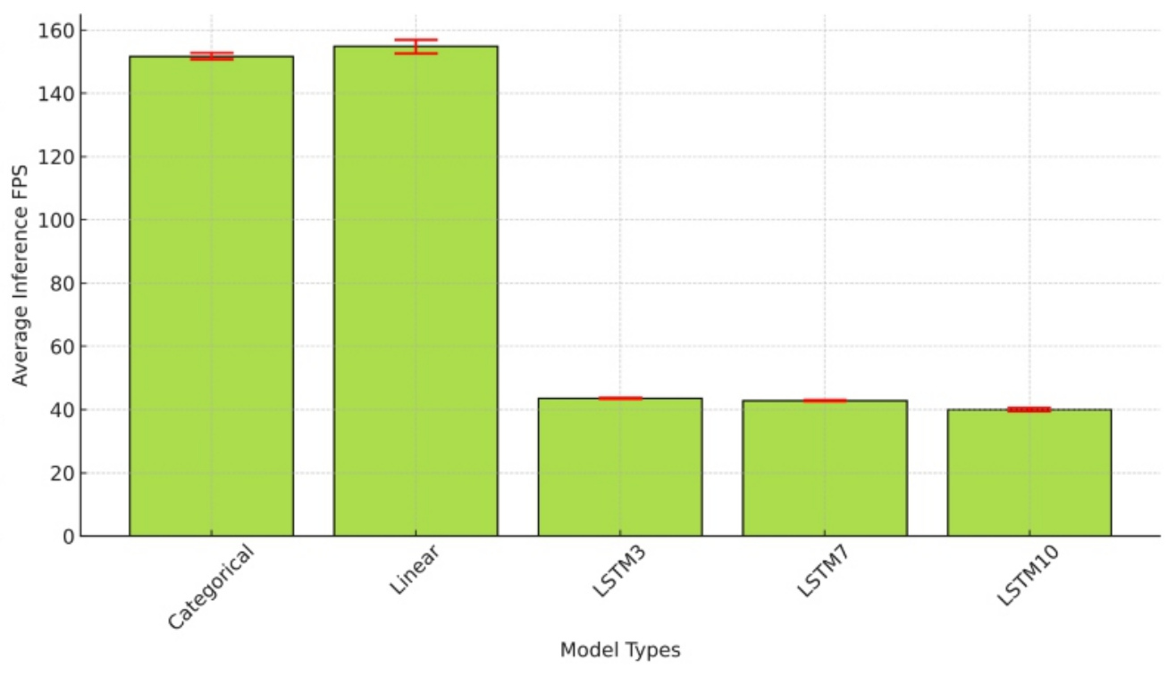

Fig. 4 Average Inference FPS for Cloud Models

Evaluation

Autonomy Calculation Formula: $\text{autonomy} = \left(1 - \frac{\text{interventions} \times 4.5\,\text{seconds}}{\text{total time}}\right) \times 100\%$

- The autonomy **scores of the cloud** aided models perform much better than those on the edge
- This is especially the case with the **LSTM models 7 and 10** that were unable to achieve any autonomy due to lack of resources

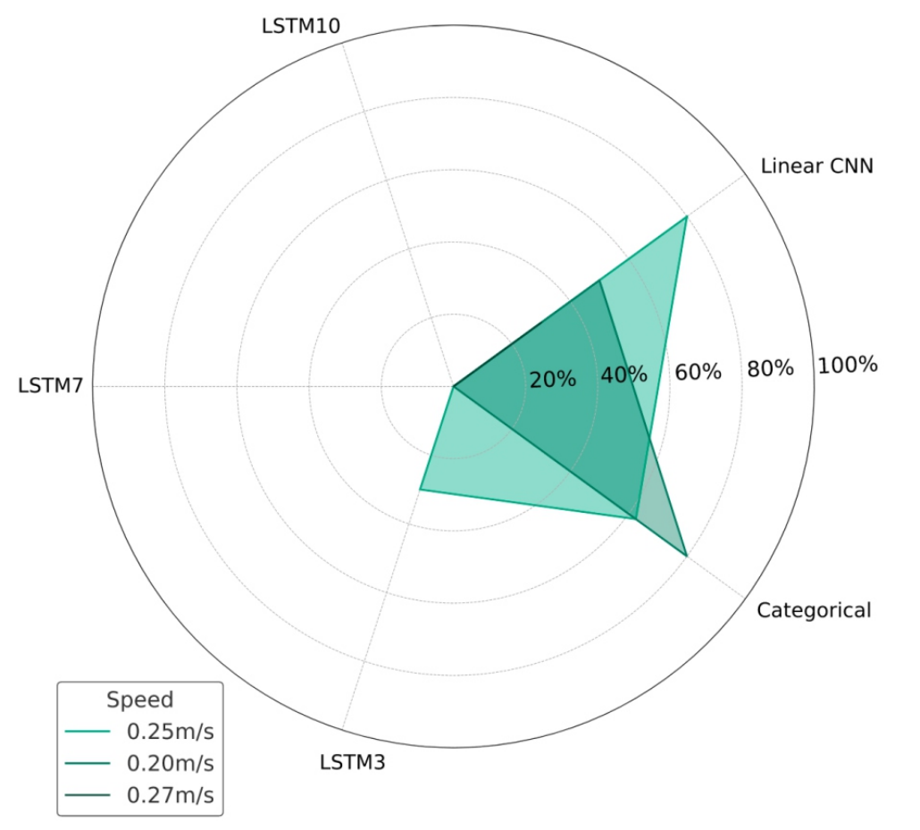

Fig. 5 Autonomy Scores of Edge Models

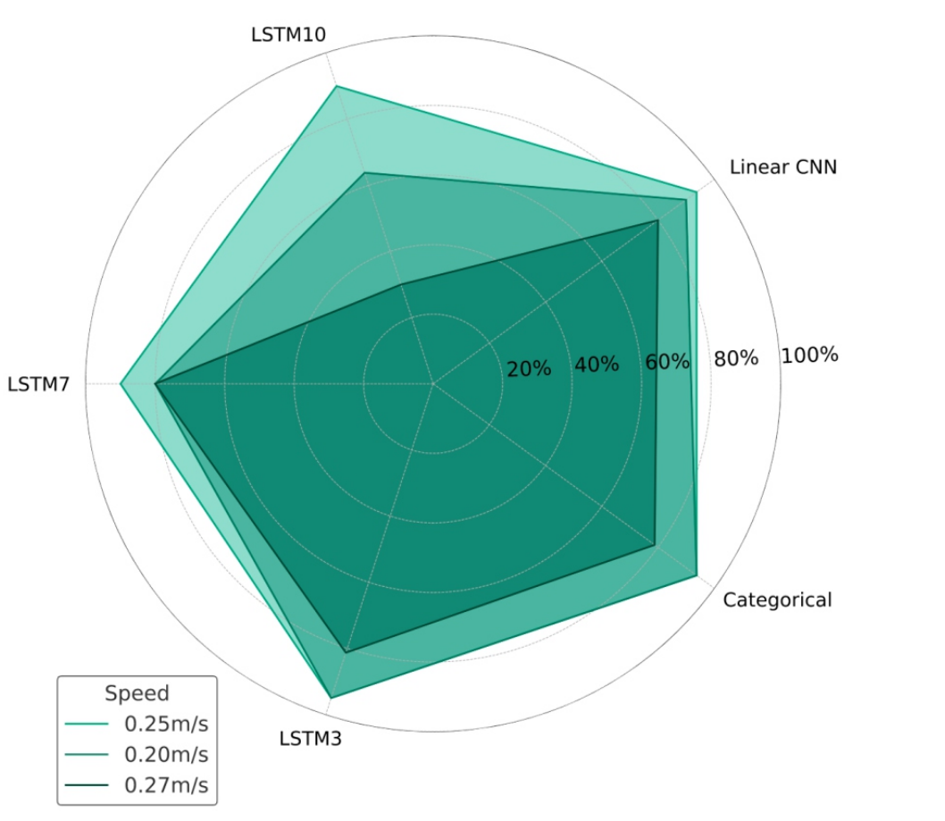

Fig. 6 Autonomy Scores of Cloud Models

- **With the cloud**, the linear model increased in autonomy by 13.75%, however, the **LSTM 10 increased by 90% at the same speeds**

- The resource usage of the **RPi4 is much lower** when the inference is offloaded to the cloud
- This is especially the case during the **LSTM3 model that uses only 50% of the CPU**, instead of **80% during pure edge operations**

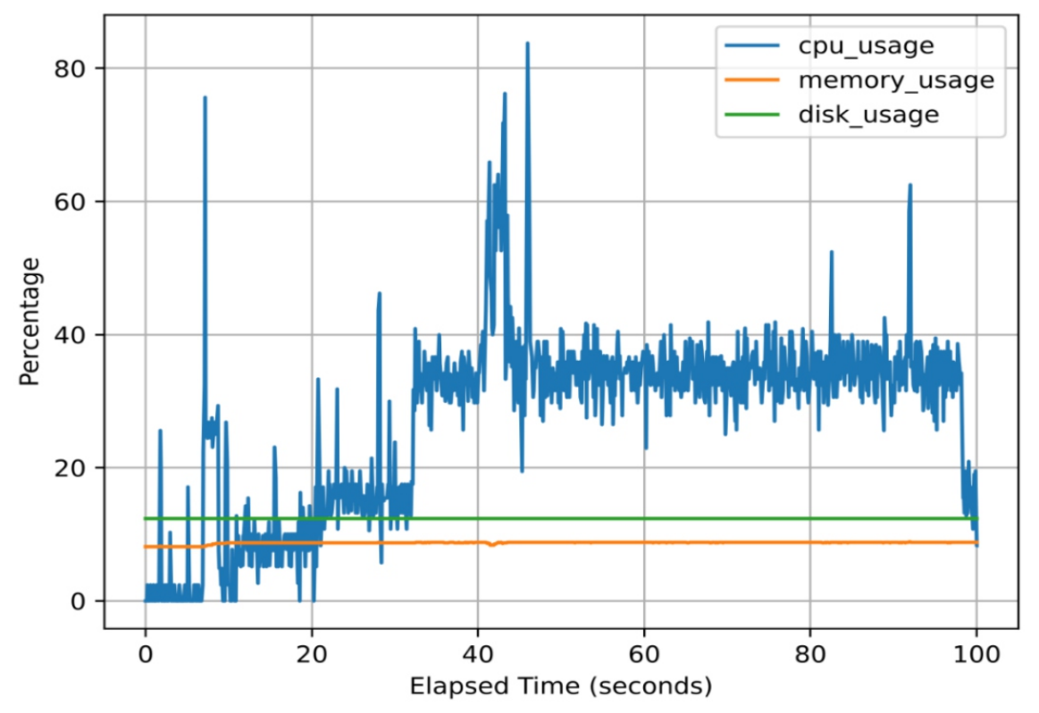

Fig. 7 Offloaded Linear Model Resource Utilization

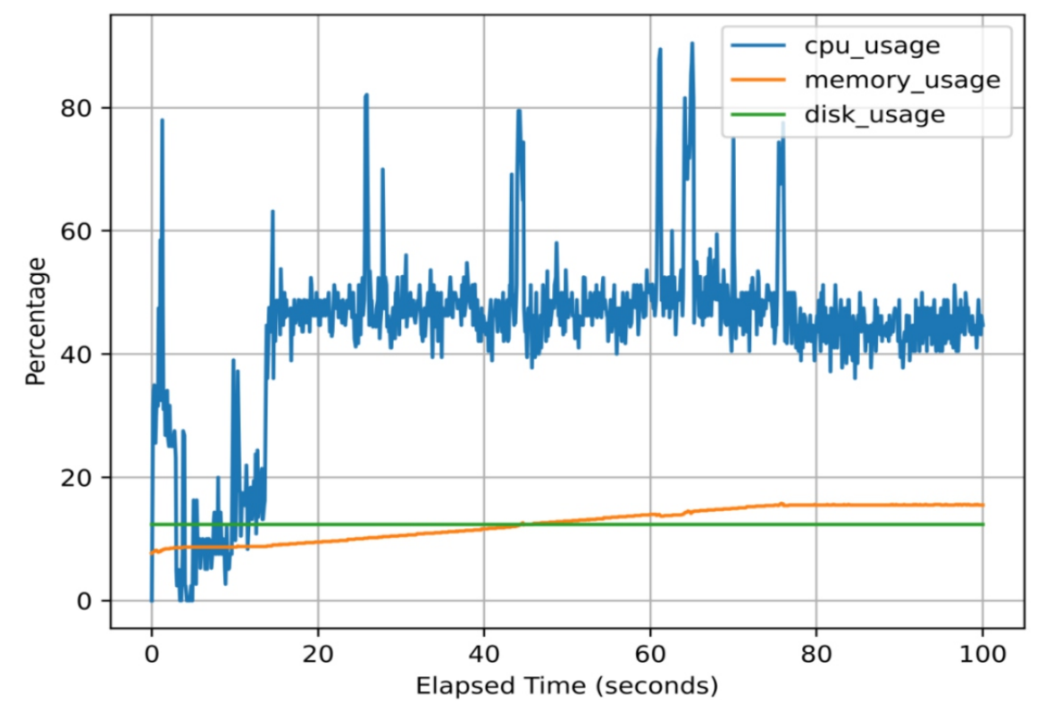

Fig. 8 Offloaded LSTM3 Model Resource Utilization

Conclusion and Future work

- **The Cloud-aided Self-Driving framework** allowed self-driving cars to offload computational load due inferencing to the cloud
- Compared to the **pure-edge framework, with the Cloud-Aided framework**, a substantial increase in terms of autonomy, especially for the LSTM models
- The **possible utility of RNNs/LSTMs** were unveiled once additional computational resources were available, **performing as well as or better than the CNNs** tested in terms of autonomy

- **Future work** will focus on increasing domain adaptability and fully-leveraging cloud capabilities