WAYVE

The Frontiers of Embodied Artificial Intelligence

Cameron Tuckerman-Lee
Tech Lead, Embodied Foundation Model, Wayve

Ohio State University, 2024-11-20

Cameron Tuckerman-Lee

Tech Lead, Foundation Models, Wayve

twitter.com/tuckerman
cameron@ctuck.com

Agenda

Plan for this evening

- History of AI + Robotics
- Deep Learning and Computer Vision
- Autonomous Driving, E2E Driving
- Research at Wayve
- Parting Thoughts
- Q&A

AI + Robotics

Early Beginnings

AI + Robotics

- Imitation Game (aka the "Turing Test") coined by Alan Turing in 1950
- Dartmouth Workshop organized by McCarthy, Minsky, Rochester, and Shannon in 1956
- Unimate, the first industrial robot, was built in 1961
- Shakey the Robot, the first general purpose robot, is developed at SRI from 1966–1972

A PROPOSAL FOR THE DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE

J. McCarthy, Dartmouth College
M. L. Minsky, Harvard University
N. Rochester, I.B.M. Corporation
C.E. Shannon, Bell Telephone Laboratories

August 31, 1955

We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

The following are some aspects of the artificial intelligence problem: 1

Automatic Computers

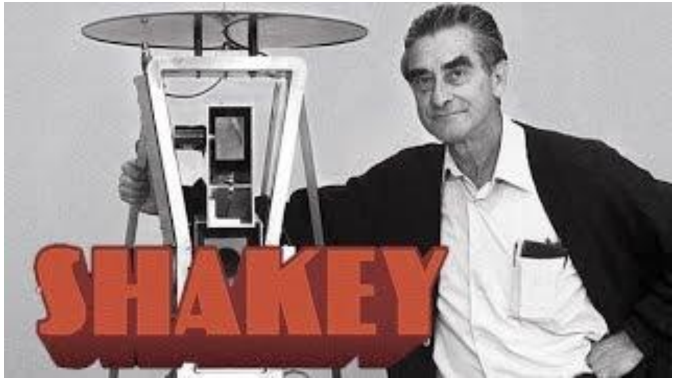
If a machine can do a job, then an automatic calculator can be programmed to simulate the machine. The speeds and memory capacities of present computers may be insufficient to simulate many of the higher functions of the human brain, but the major obstacle is not lack of machine capacity, but our inability to write programs taking full advantage of what we have. 2

Source: McCarthy, J., et al. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence.

Milestones in Robotics

AI + Robotics

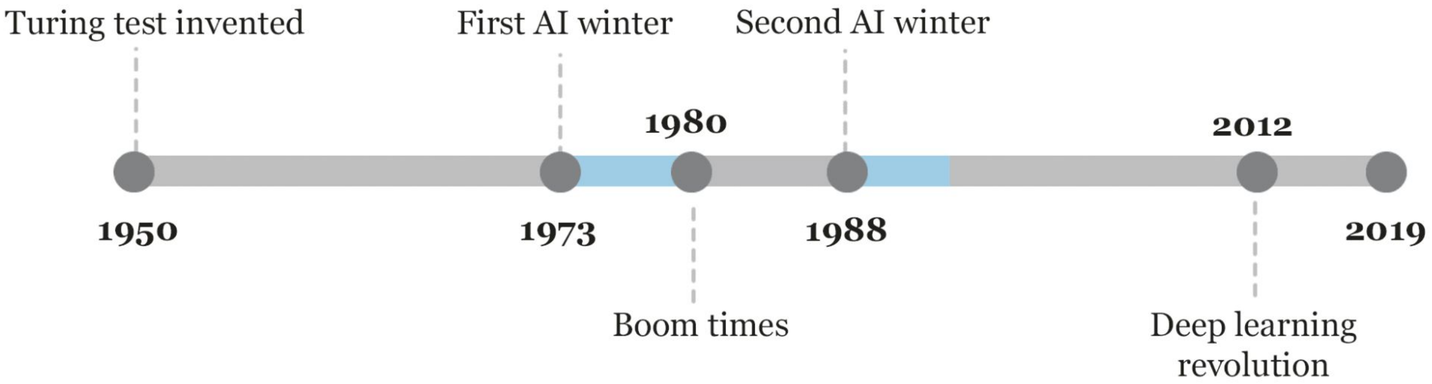
- 1966s–1972s: Shakey (first general-purpose robot)
- 1980s–90s: Autonomous vehicles (DARPA-funded, like ALVINN)
- Late 90s, 2000s: Introduction of ASIMO by Honda, Boston Dynamics



AI Winter(s)

AI + Robotics

- First winter in 1974–1980, second winter (the one you might be more familiar with) was 1987-2000.
- Causes: Unrealistic expectations, limited computing power
- Impact: Funding cuts, skepticism among researchers

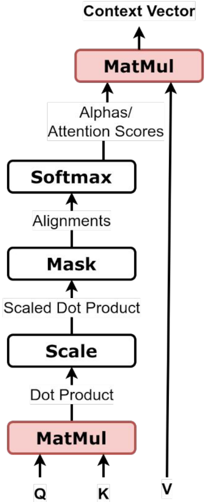


Source: Schuchmann, S. (2019). History of the first AI Winter.

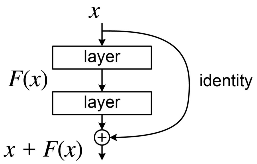
Resurgence of AI: AI Spring 2012

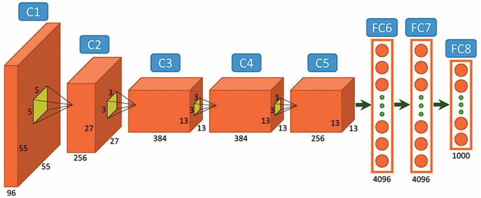
AI + Robotics

- Key events: ImageNet (and AlexNet), advances in deep learning
- Moore's Law: Driving the computational power necessary for breakthroughs



Scaled Dot-Product
Attention





AI has surpassed humans at a number of tasks and the rate at which humans are being surpassed at new tasks is increasing

State-of-the-art AI performance on benchmarks, relative to human performance

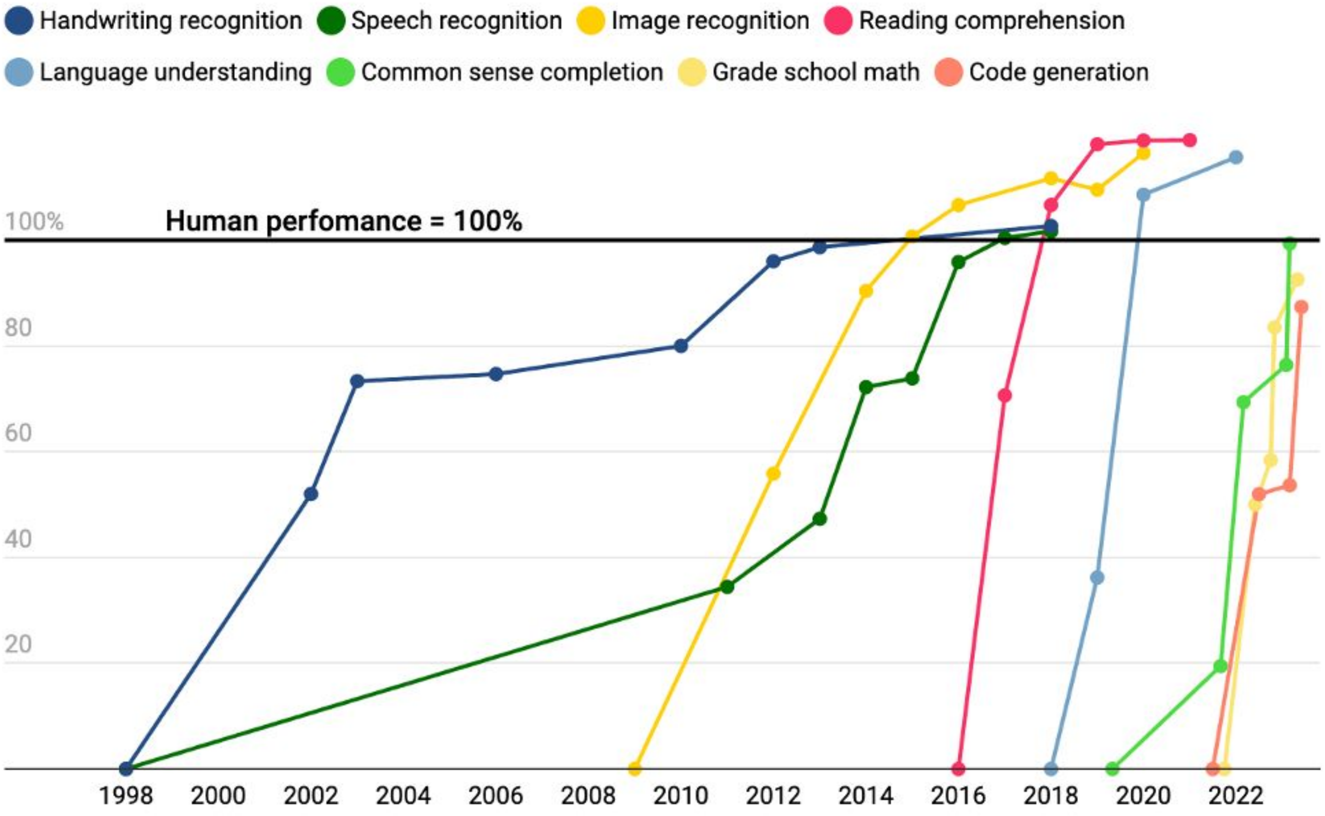
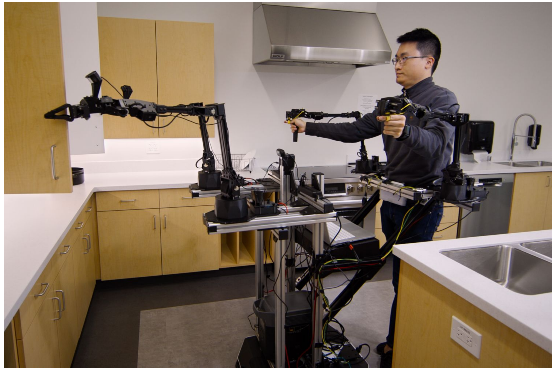


Chart: Will Henshall for TIME • Source: ContextualAI

Embodied AI

AI + Robotics

- Systems that perceive, interact, and learn from the physical world
- Examples: Robotics in manufacturing, smart home assistants, self-driving cars

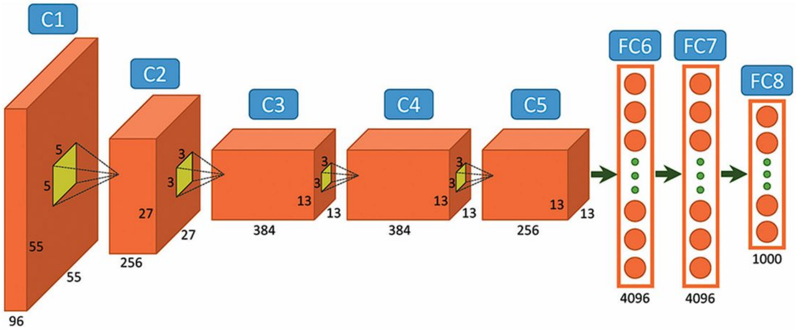


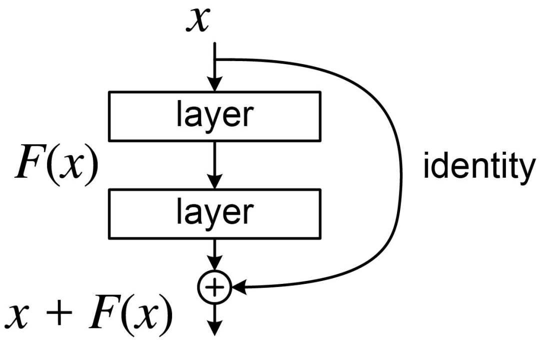

Deep Learning
+
Computer Vision

ImageNet Moment

Deep Learning + Computer Vision

- Database launches at CVPR 2009, competition launches in 2010
- 2010-2012 dominated by classical models (e.g. SVMs)
- 2012: AlexNet winning ImageNet competition, sparking renewed interest in neural networks (and GPU training!)
- 2015: Superseded by “Very Deep CNNs” or ResNets





Object Detection, Segmentation

Deep Learning + Computer Vision

- Early example: SegNet (**Vijay Badrinarayanan**, **Alex Kendall**, Roberto Cipolla)
- Key models: YOLO, Faster R-CNN, Mask R-CNN, DETR
- Applications: Self-driving cars, healthcare, industrial automation

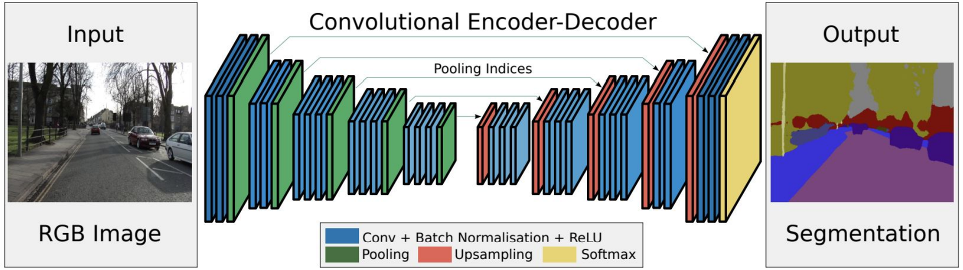
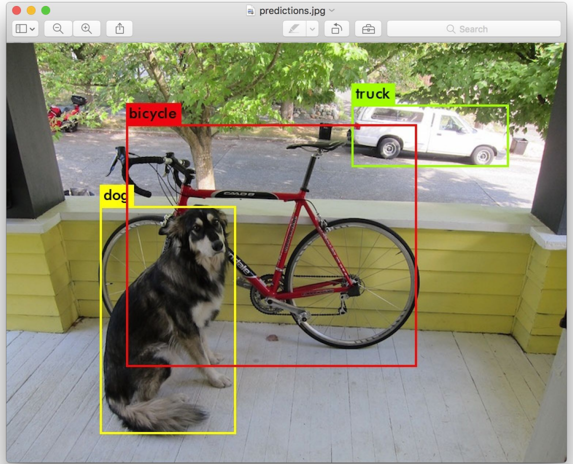


Fig. 2. An illustration of the SegNet architecture. There are no fully connected layers and hence it is only convolutional. A decoder upsamples its input using the transferred pool indices from its encoder to produce a sparse feature map(s). It then performs convolution with a trainable filter bank to densify the feature map. The final decoder output feature maps are fed to a soft-max classifier for pixel-wise classification.

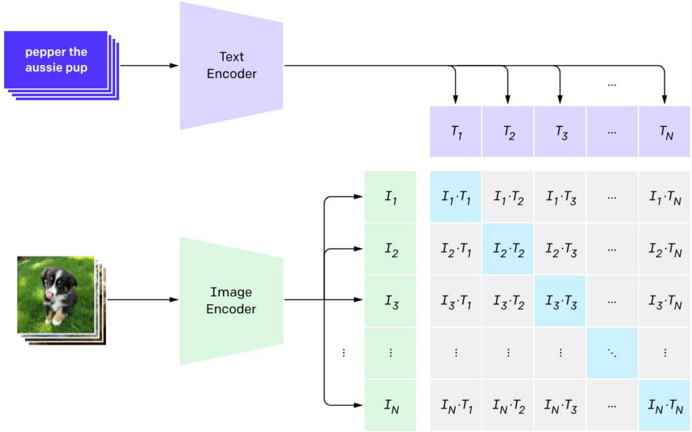


Source: Schuchmann, S. (2019). History of the first AI Winter.

Transfer Learning

Deep Learning + Computer Vision

- Using pre-trained models for new tasks (CLIP, DINO)
- Reducing training times, increasing accessibility of AI tools

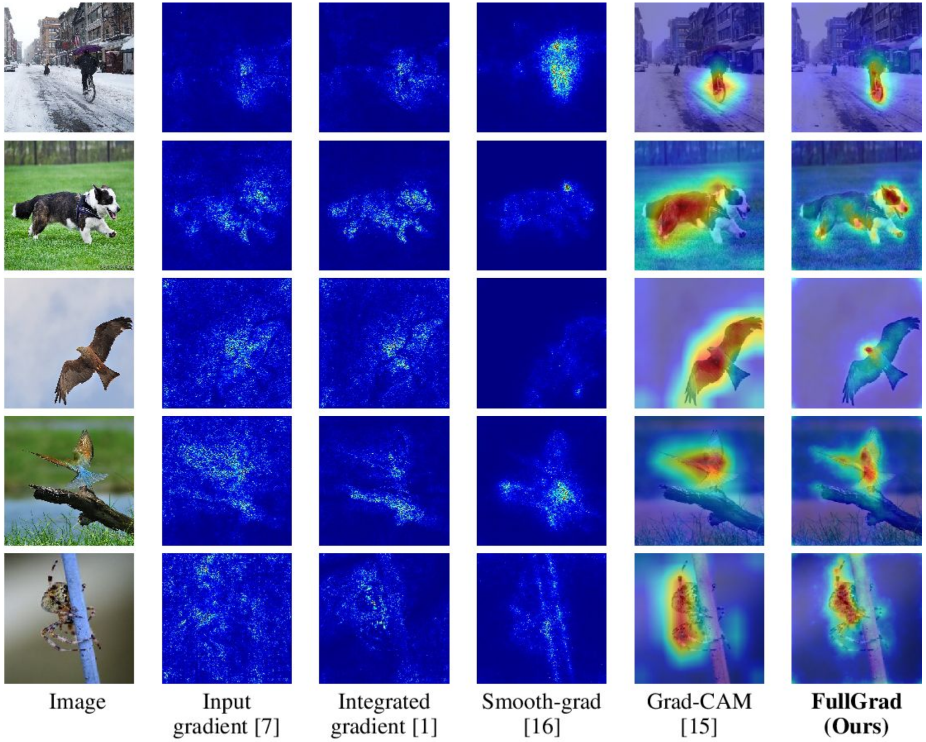


Source: Radford, A., et al. (2021). Learning Transferable Visual Models From Natural Language Supervision.
Source: Caron, M., et al. (2021) Emerging Properties in Self-Supervised Vision Transformers.

Explainability

Deep Learning + Computer Vision

- Safety-critical applications like healthcare and autonomous driving require explainability
- Methods: Saliency maps, feature visualization

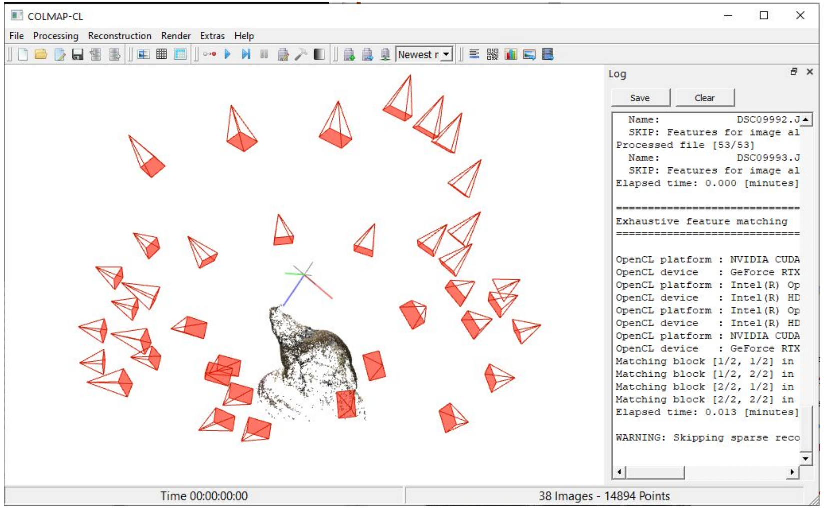


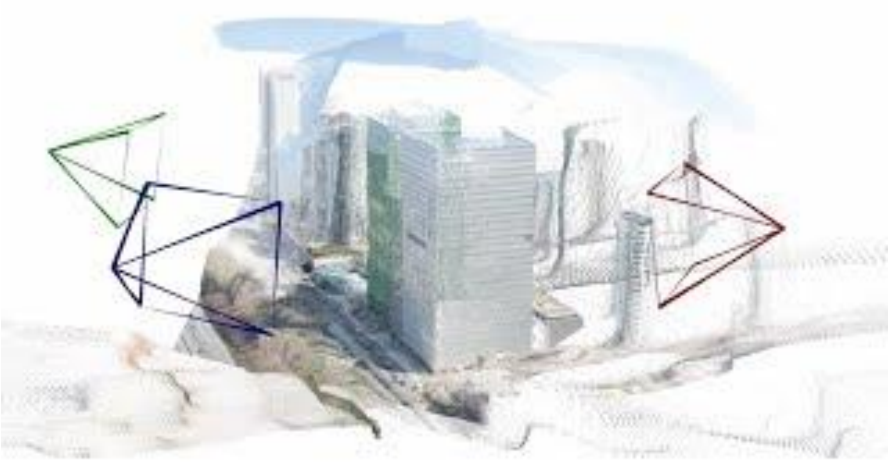
Source: Srinivas, S., Fleuret, Francois. (2019). Full-Gradient Representation for Neural Network Visualization.

3D (and 4D)

Deep Learning + Computer Vision

- Understanding depth (3D) and time (4D) in real-world perception
- Simultaneous Localization and Mapping (SLAM), 3D object detection, etc





Source: Wang, S., et al. (2023). DUSt3R: Geometric 3D Vision Made Easy.

Autonomous Driving

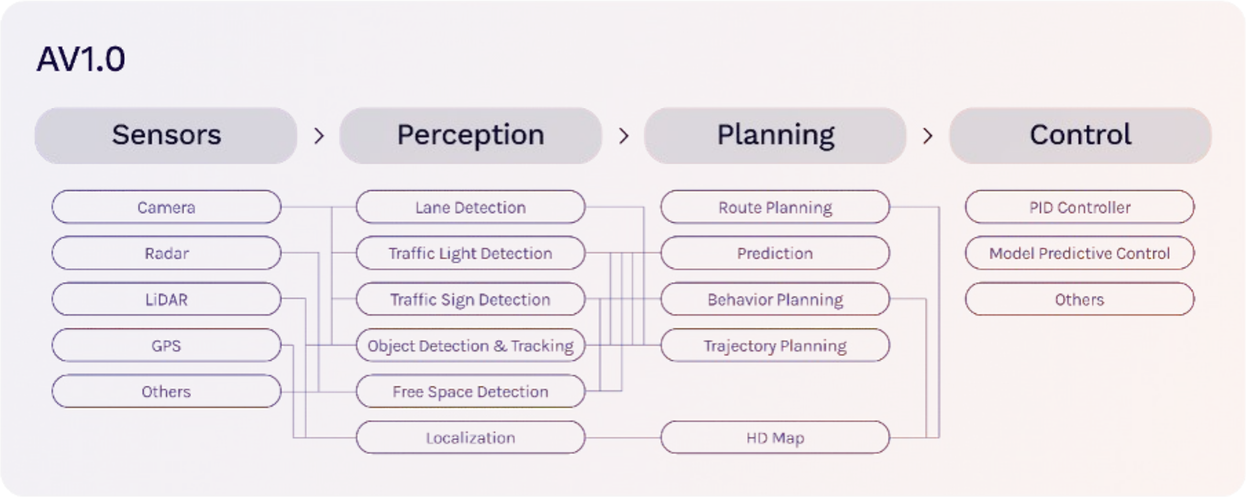
Why AV?

Autonomous Driving

- Reliable robots
- Easy to collect expert demonstrations
- Expert demonstrations come with lots of diversity, utility, volume
- Valuable service (utility, safety, efficiency, …)

The Problem

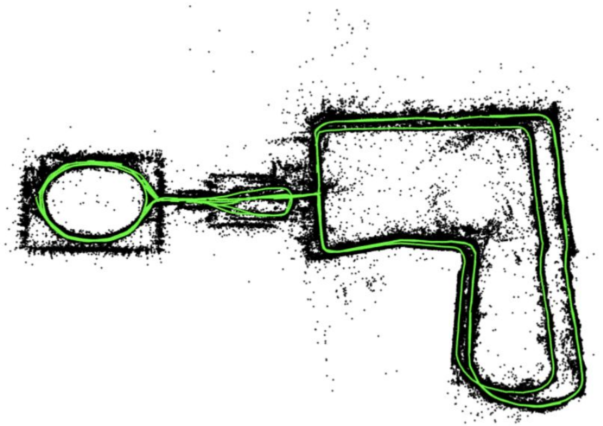
Autonomous Driving

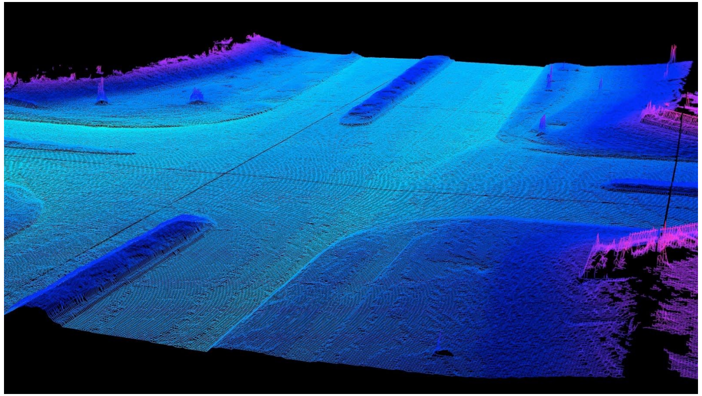


SLAM and Mapping

Autonomous Driving

- Techniques: Using sensors to build and update maps in real-time
- Importance: Localization and navigation for autonomous systems





Source: Mur-Artal, R., et al. (2015). ORB-SLAM: a Versatile and Accurate Monocular SLAM System.
Source: Waymo Team. (2016). Building maps for a self-driving car.

Challenges of Modular Approach

Autonomous Driving

- Rule-based systems: Hand-coded rules and logic
- Long-tail makes or breaks driving: complex, unpredictable environments

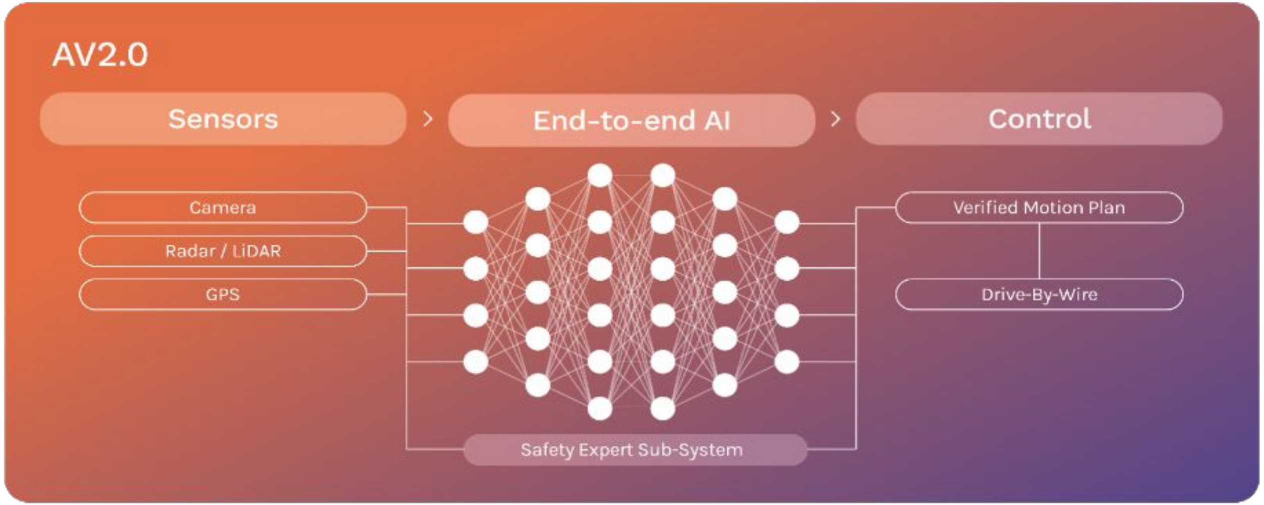
E2E Driving

AV 2.0


WING-STOP
PALACE
Harry Pot

AV2.0

E2E Driving



Advantages

E2E Driving

Wayve's Approach: Using E2E learning for real-world driving in complex cities

- Simplifying the traditional autonomous vehicle stack
- Potential for improved generalization and adaptability
- Reduced engineering complexity
- Faster adaptation to new environments

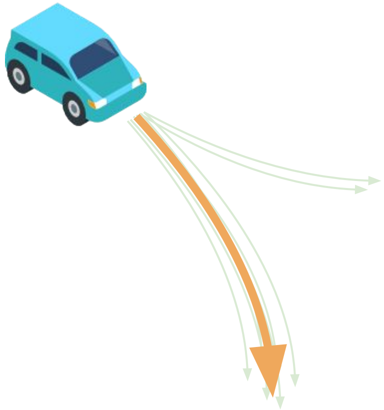
DreamWorks

HOW TO TRAIN YOUR MODEL

Approaches

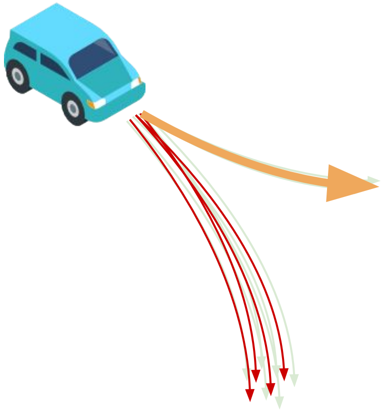
Imitation Learning (IL)

IL tries to copy the most “popular” *positive demonstration.*



Reinforcement learning (RL)

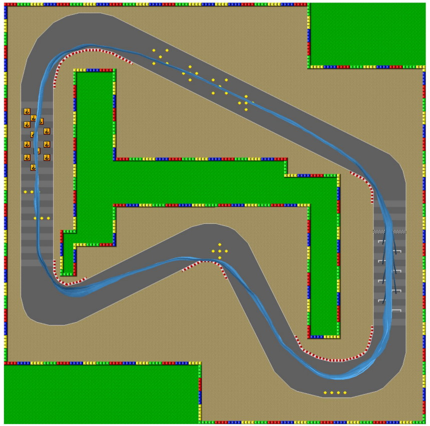
RL tries to seek *positive feedback* while avoiding *negative feedback.*



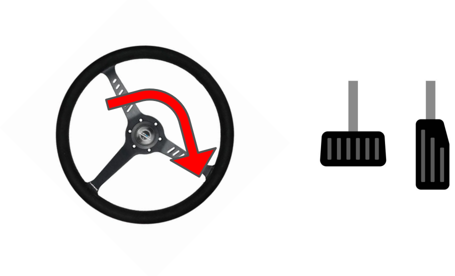
Imitation Learning

$$\hat{\pi}(s) = \underset{\pi\in\Pi}{\operatorname{argmin}} \underset{s\sim d_{\pi^*}}{\mathbb{E}} [(\ell(s,\pi))]$$

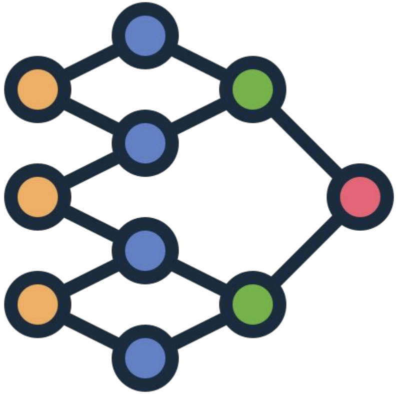
Expert Demonstrations



State/Action Pairs

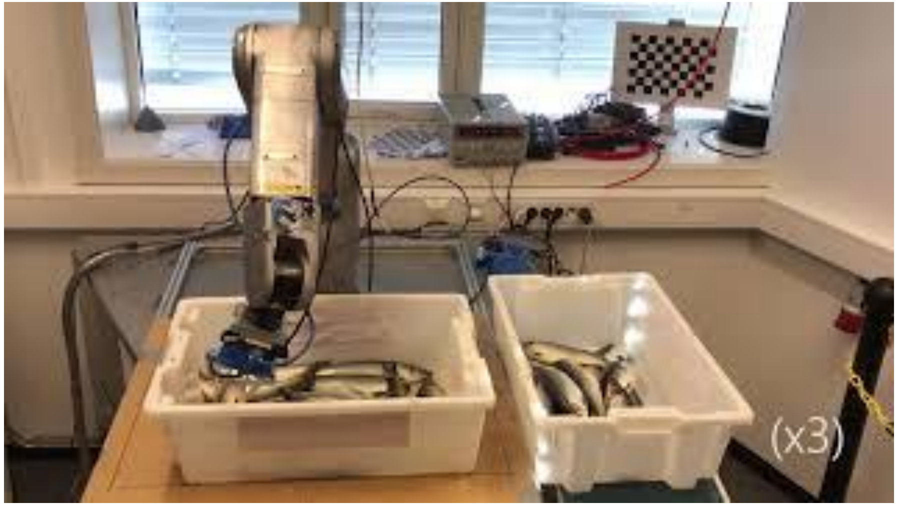


Learning



Imitation Learning

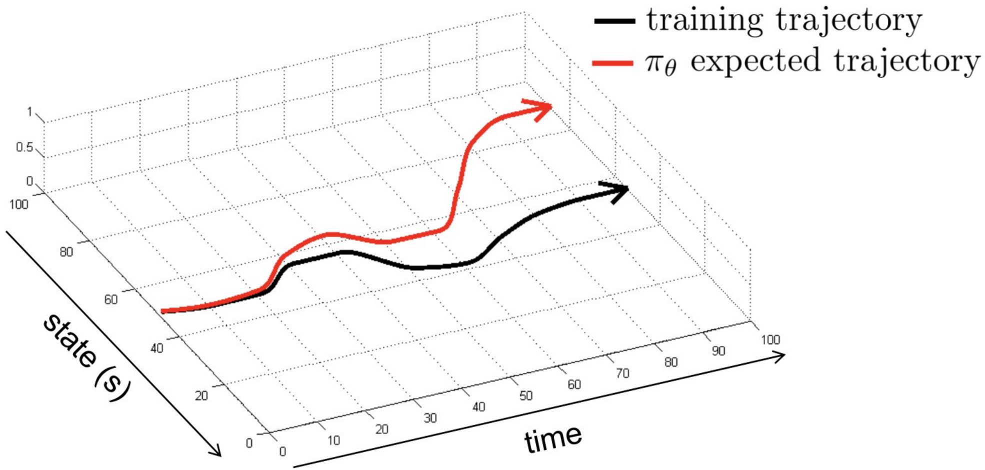
Does it work?



Accumulating errors in IL

Imitation Learning

Imitation Learning often encounters accumulating errors when deployed in practice due to **distribution shift**: the data-collecting policy differs from the learned policy



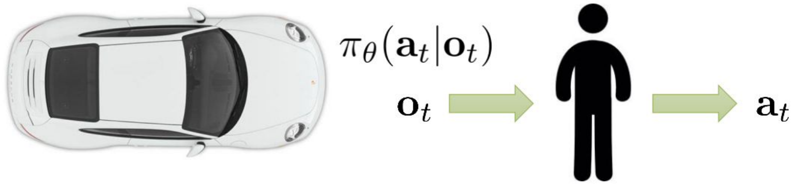
Source: Levine, S., et al. (2020). Deep Reinforcement Learning.

DAgger

Imitation Learning

Idea: Solve distribution shift by collecting expert demonstrations on-policy!

1. train $\pi_\theta(\mathbf{a}_t|\mathbf{o}_t)$ from human data $\mathcal{D} = \{\mathbf{o}_1, \mathbf{a}_1, \ldots, \mathbf{o}_N, \mathbf{a}_N\}$
2. run $\pi_\theta(\mathbf{a}_t|\mathbf{o}_t)$ to get dataset $\mathcal{D}_\pi = \{\mathbf{o}_1, \ldots, \mathbf{o}_M\}$
3. $\boxed{\text{Ask human to label } \mathcal{D}_\pi \text{ with actions } \mathbf{a}_t}$
4. Aggregate: $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_\pi$

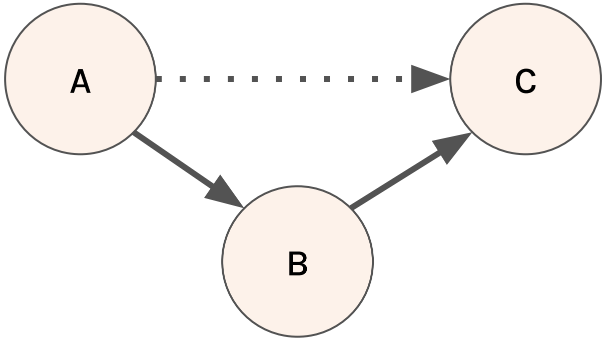


Source: Ross, S., et al. (2010). A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning
Source: Levine, S., et al. (2020). Deep Reinforcement Learning.

IL limitations

Imitation learning

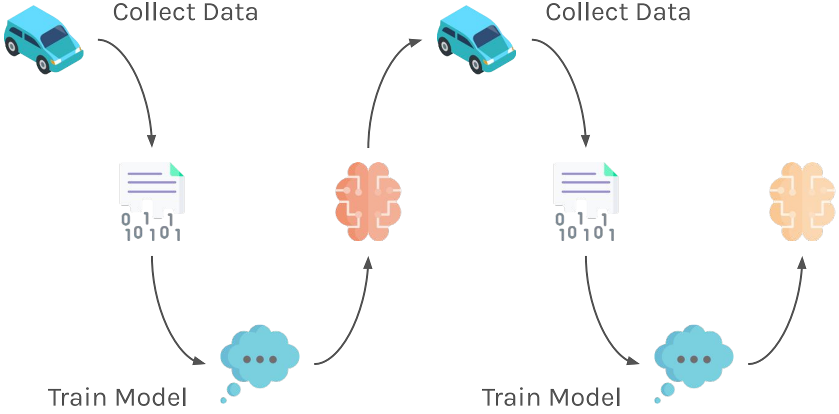
- IL cannot perform better than human expert. DAgger is expensive in the real world!
- Can we learn from non-expert data and perform better than human experts?
  - Can we learn about the action that directly takes us from A to C?



Types of RL Algorithms

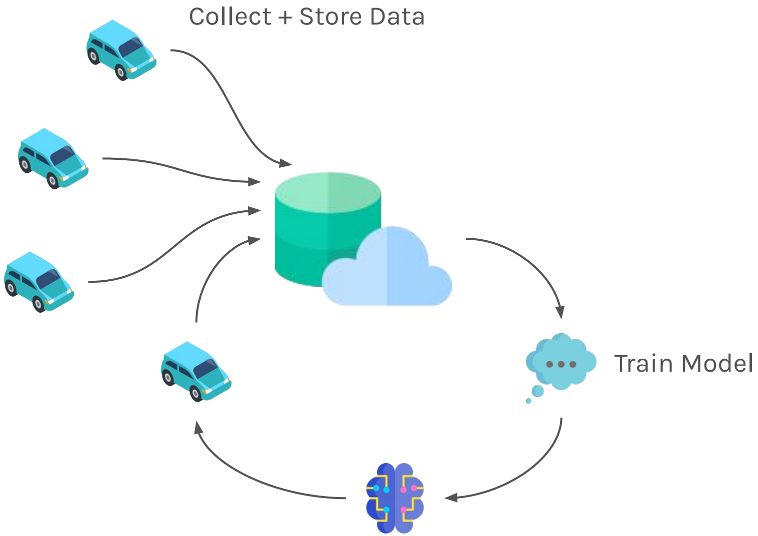
On-policy

Improve the policy with data collected by the *current* policy.



Off-policy

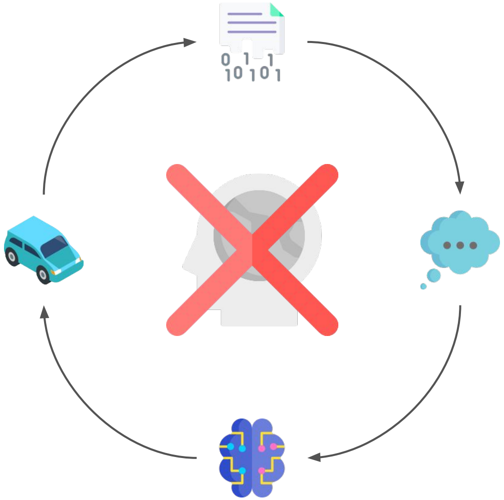
Improve the policy with data collected by *any* policy.



Types of RL Algorithms

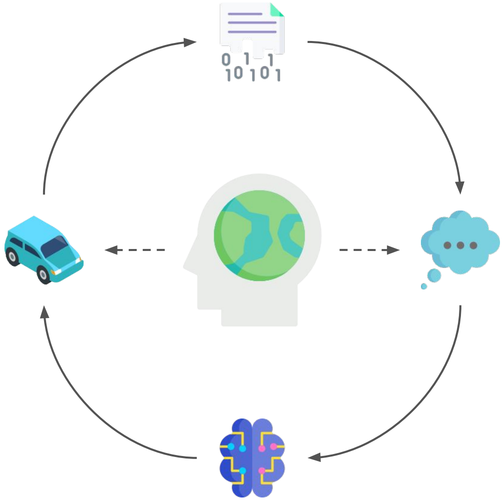
Model-free

Training (and inference) have access to *experience* only.

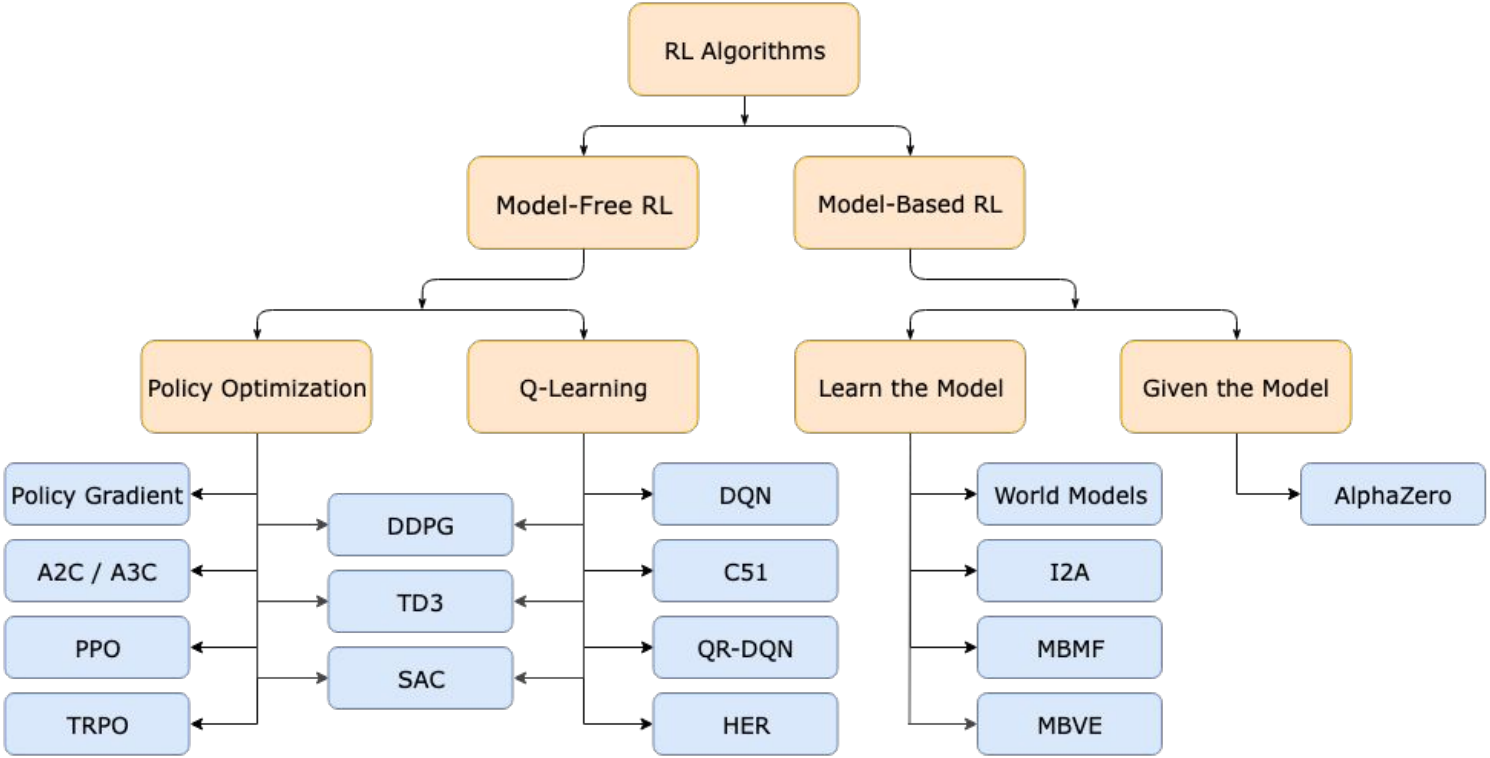


Model-based

Training (and inference) have access to both experience and a *world model*.



Types of RL Algorithms



Source: Achiam, J. Spinning Up.

Q-Learning

Reinforcement Learning

Off-policy RL algorithm that estimates expected future rewards (Q-value) given an action-state pair

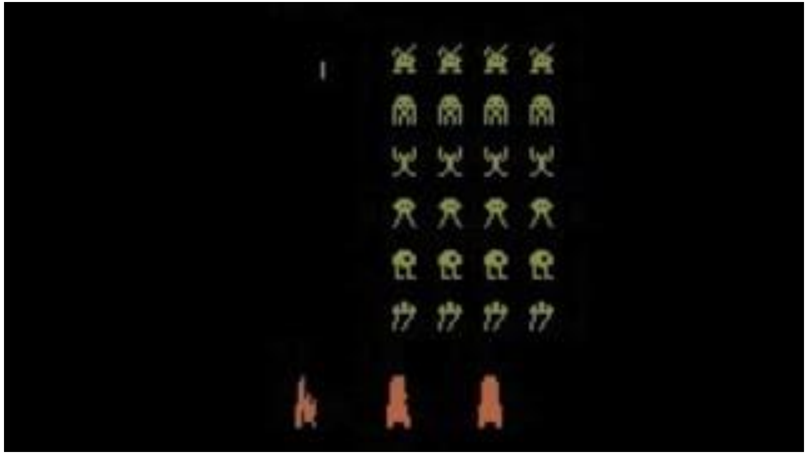
Updates are made using the Bellman equation:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left[R(s,a) + \gamma \max_{a'} Q(s',a') - Q(s,a) \right]$$

DQL

Reinforcement Learning

Use a neural network to learn the Q function!

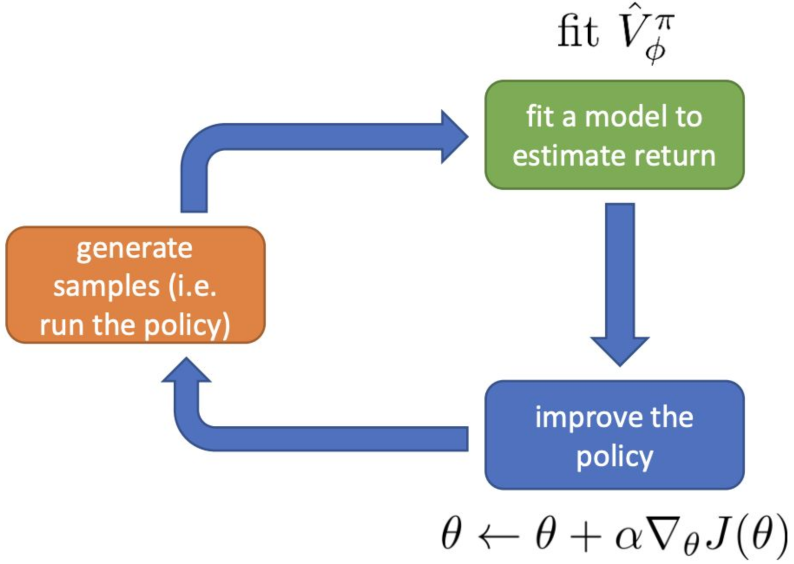


Actor Critic

Reinforcement Learning

- **Actor:** learns to maximize performance under critic
  - Trained with policy gradients
- **Critic:** learns to estimate action-values
  - Trained with Monte Carlo / TD updates

Basic recipe covers all modern model-free RL algorithms (e.g., SAC, PPO, TD3)



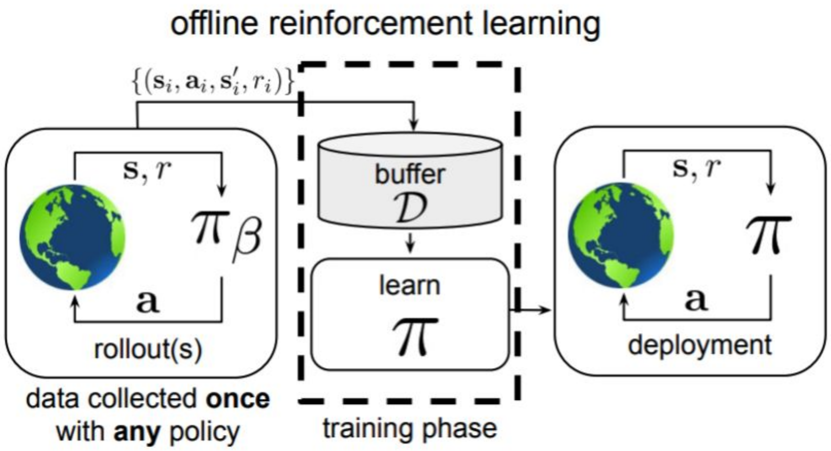
Source: Levine, S., et al. (2020). Deep Reinforcement Learning.

Offline RL

Reinforcement Learning

Uses pre-collected data to train a policy without needing to directly interact with the environment

Can greatly improve scale as data collection is cheaper than deployment

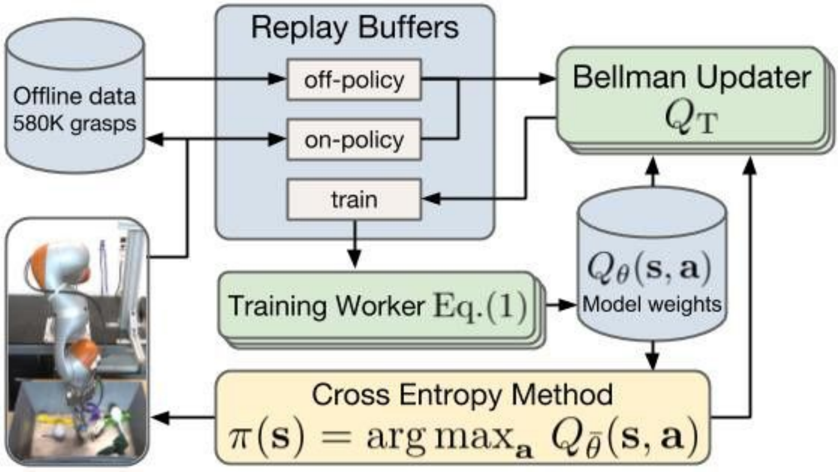


Source: Levine, S., et al. (2020). Deep Reinforcement Learning.

Mixed Online/Offline RL

Reinforcement Learning

Some approaches provide for mixing on-policy and off-policy data



Source: Kalashnikov, D., et al. (2018) QT-Opt: Scalable Deep Reinforcement Learning for Vision-Based Robotic Manipulation

Research at Wayve

History

Research @ Wayve

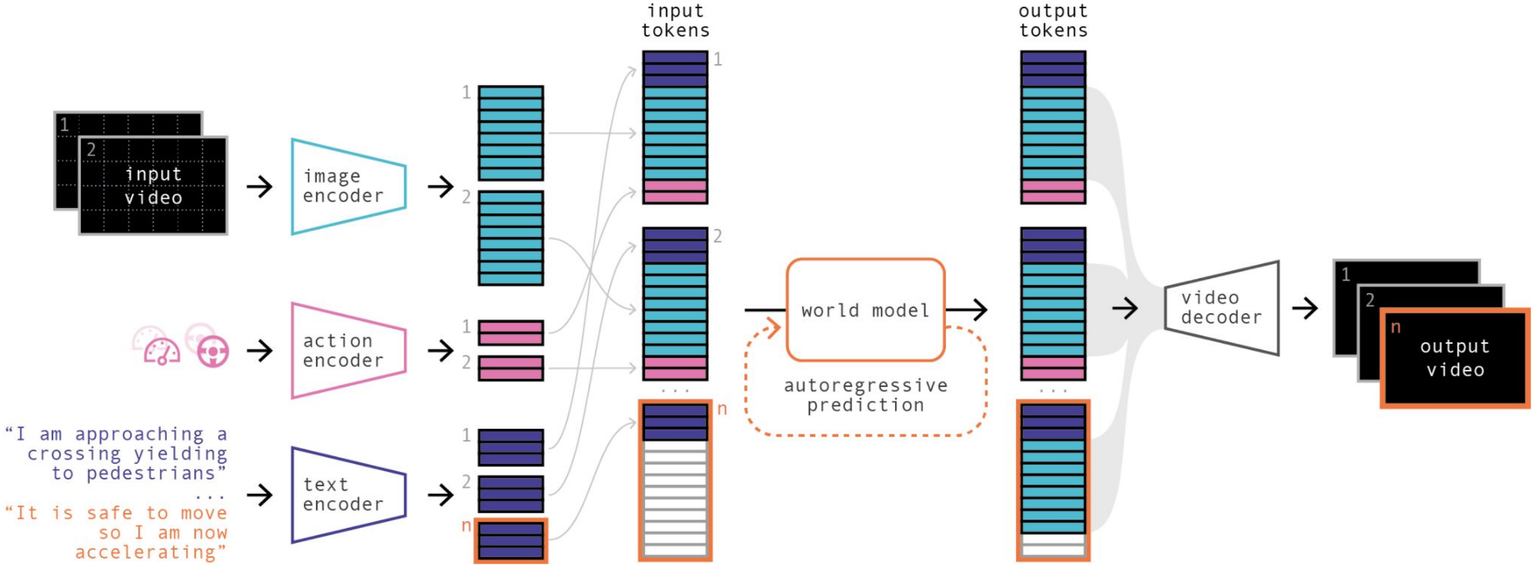
- 2017: Founded in Cambridge
- 2019: Wayve was the first company to demonstrate an end-to-end learned driving system on UK public roads
- 2022: Demonstrated AI model driving multiple types of vehicles and in multiple cities across the UK
- 2023-Present: …


WAYVE
4X SPEED

GAIA

World Modelling

GAIA-1



Hu, A., et al. (2023). GAIA-1: A Generative World Model for Autonomous Driving. http://arxiv.org/abs/2309.17080

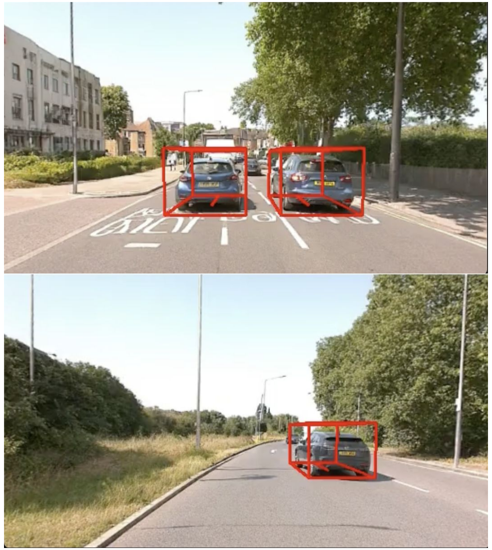

4X SPEED


Generated by GAIA-1
Generated by GAIA-1
Generated by GAIA-1
Generated by GAIA-1

Controlling Dynamic Agents

GAIA-1

Extension to GAIA-1 demonstrated at CVPR 2024...



Controlling Dynamic Agents

GAIA-1

Extension to GAIA-1 demonstrated at CVPR 2024...

PRISM + Ghost Gym

Scene Reconstruction

Ghost Gym & PRISM

Zürn, J., et al. (2024). WayveScenes101: A Dataset and Benchmark for Novel View Synthesis in Autonomous Driving. https://arxiv.org/abs/2407.08280
Introducing PRISM-1: Photorealistic reconstruction in static and dynamic scenes. https://wayve.ai/thinking/prism-1/

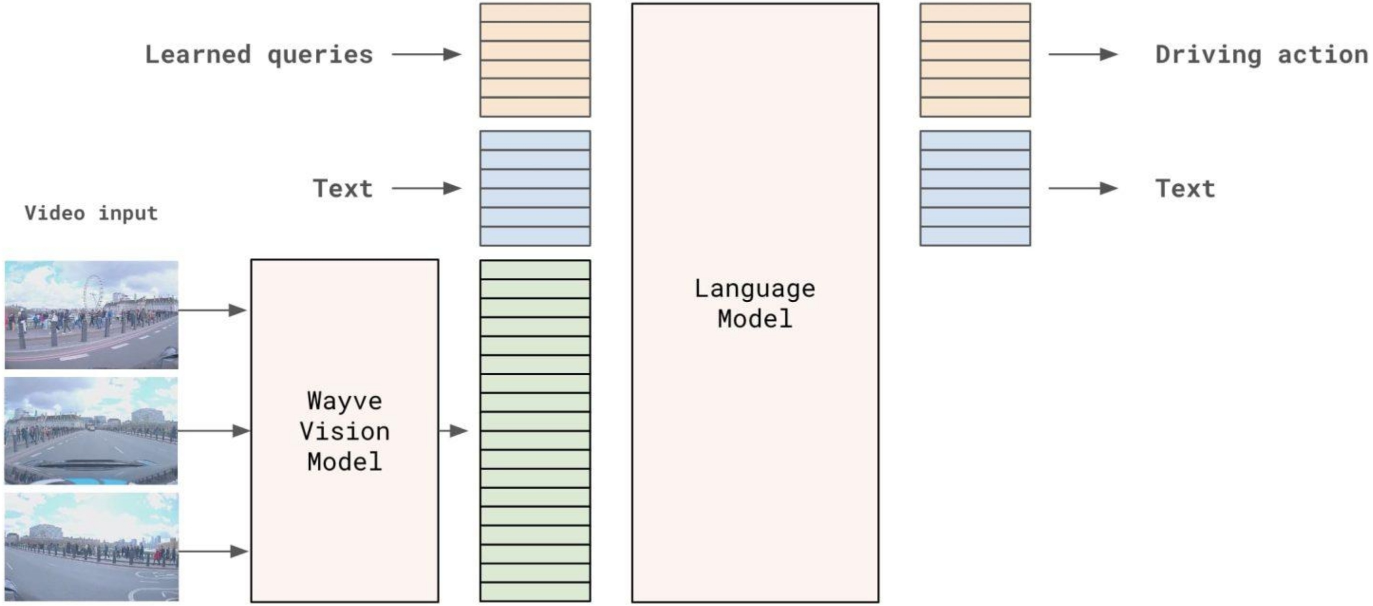

Freeze Time


WAYVE
Freeze Time

LINGO

Language

LINGO-1 & LINGO-2



Source: Marcu, A., et al. (2023). LingoQA: Visual Question Answering for Autonomous Driving.


WAYVE
Select another route
London N Route


WAYVE


Autonomous: True
Speed: 13.98 mph
Keeping pace with the lead vehicle.

Counterfactuals (GAIA + LINGO)

LINGO

Foundation Models

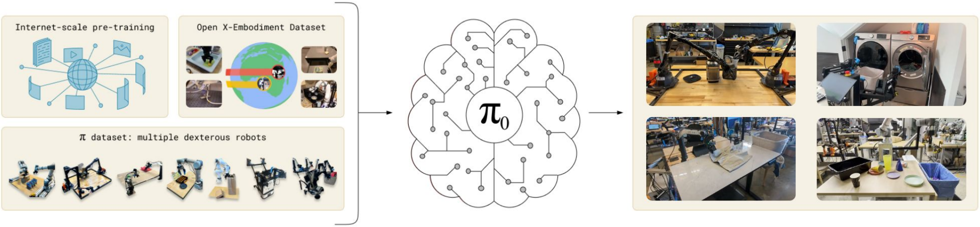
Foundation Models Today

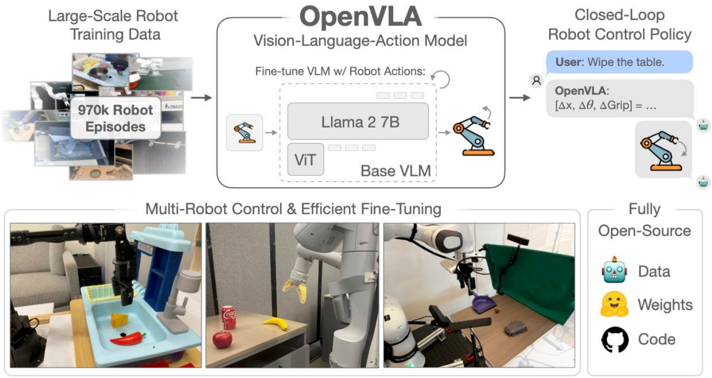
Foundation Models

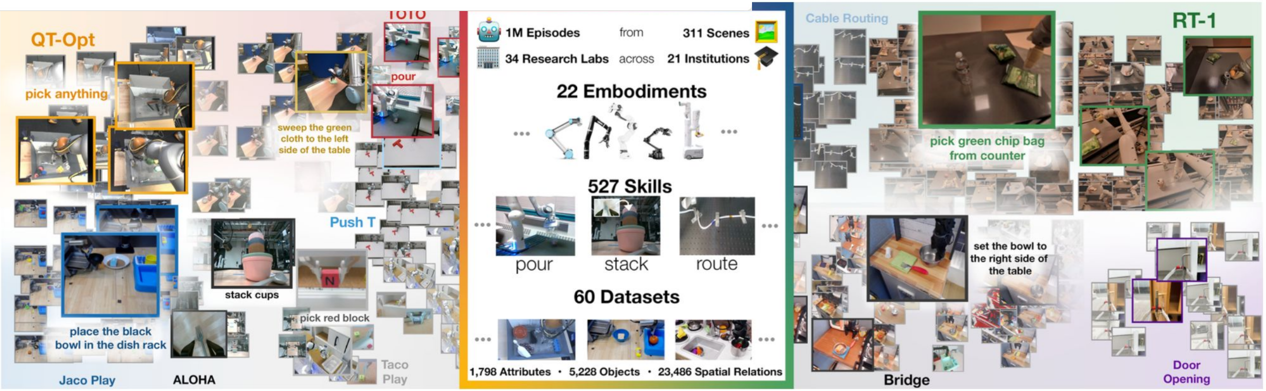
- Examples
  - Language: GPT, Gemini, Llama
  - Vision: CLIP, DINO, JEPA, MAE
  - Embodied: OpenVLA, π0
- Challenges
  - Many robotics tasks need great video features that include 3D/4D understanding!
  - Many robotics tasks require inference at > XX Hz frequencies!

FMs ⇔ Robotics

Foundation Models







Black, K., et al. (2024). π0: A Vision-Language-Action Flow Model for General Robot Control.
Kim, M., et al. (2024). OpenVLA: An Open-Source Vision-Language-Action Model.
Open X-Embodiment Collaboration, et al. (2023). Open X-Embodiment: Robotic Learning Datasets and RT-X Models.

Parting Thoughts

Takeaways

Parting Thoughts

- Simple objectives perform surprisingly well when trained at scale (in terms of both model size and data corpus size)
- The trend across robotics is moving from modular systems to end-to-end systems, including our simulators
- Foundation models are increasingly becoming critical components of embodied systems
- Diversity, quantity, and quality of data is **key** to solving robotics

WAYVE

References

McCarthy, J., et al. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence.

Schuchmann, S. (2019). History of the first AI Winter.

Badrinarayanan, V., Kendall, A., Cipolla, R. (2015). SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation.

Radford, A., et al. (2021). Learning Transferable Visual Models From Natural Language Supervision.

Caron, M., et al. (2021) Emerging Properties in Self-Supervised Vision Transformers.

Srinivas, S., Fleuret, Francois. (2019). Full-Gradient Representation for Neural Network Visualization.

Wang, S., et al. (2023). DUSt3R: Geometric 3D Vision Made Easy.

Mur-Artal, R., et al. (2015). ORB-SLAM: a Versatile and Accurate Monocular SLAM System.

Waymo Team. (2016). Building maps for a self-driving car.

Levine, S., et al. (2020). Deep Reinforcement Learning.

Ross, S., et al. (2010). A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning

Achiam, J. Spinning Up.

Levine, S., et al. (2020). Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems.

Mnih, V., et al. (2013). Playing Atari with Deep Reinforcement Learning.

Kalashnikov, D., et al. (2018) QT-Opt: Scalable Deep Reinforcement Learning for Vision-Based Robotic Manipulation

Hu, A., et al. (2023). GAIA-1: A Generative World Model for Autonomous Driving.

Introducing PRISM-1: Photorealistic reconstruction in static and dynamic scenes.

Zürn, J., et al. (2024). WayveScenes101: A Dataset and Benchmark for Novel View Synthesis in Autonomous Driving.

Marcu, A., et al. (2023). LingoQA: Visual Question Answering for Autonomous Driving.

Black, K., et al. (2024). π0: A Vision-Language-Action Flow Model for General Robot Control.

Kim, M., et al. (2024). OpenVLA: An Open-Source Vision-Language-Action Model.

Open X-Embodiment Collaboration, et al. (2023). Open X-Embodiment: Robotic Learning Datasets and RT-X Models.