Reusable (Robot) Learning

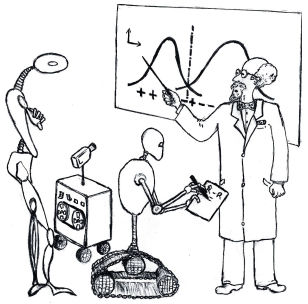
Markus Wulfmeier
Postdoctoral Research Scientist
Oxford Robotics Institute

Prev. Visiting Scholar
Berkeley AI Research
UC Berkeley

Source: Salisbury Robotics Lab, Stanford University, 2007

A Simplified View of Automation

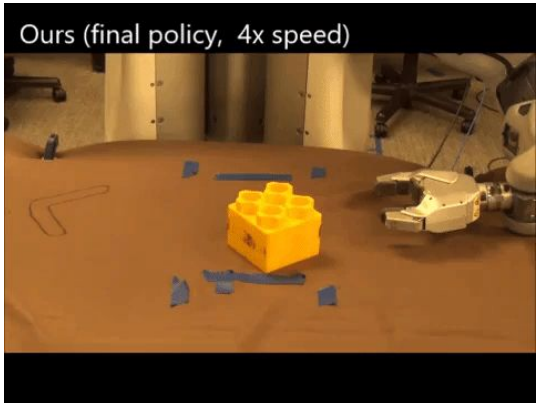
- Pre-Automation → **act** yourself
- 1. Automation → **program** rules
- 2. Automation → **learn/teach** rules

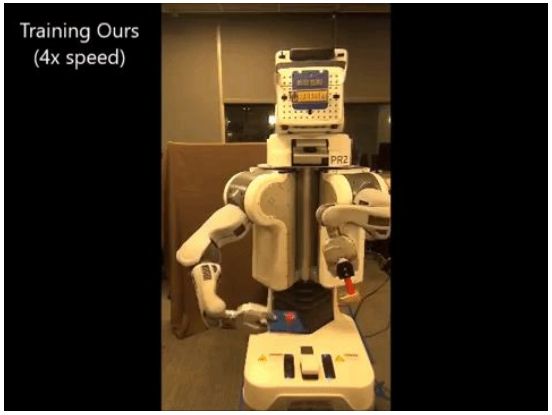


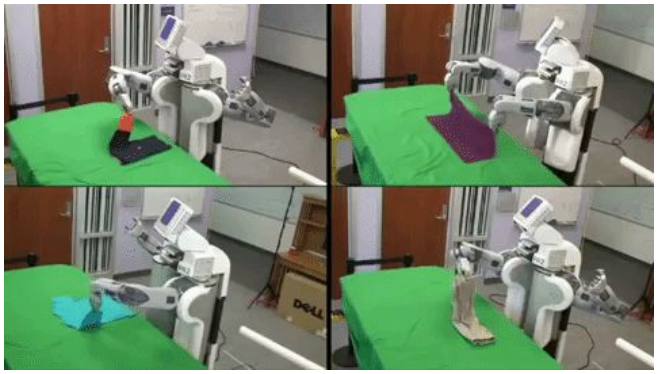
Sources: princeton.edu & pixabay.com

Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

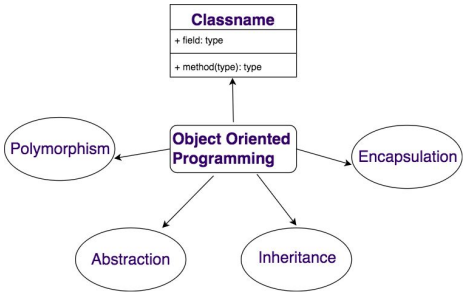


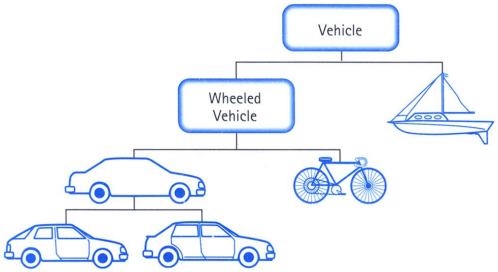




Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

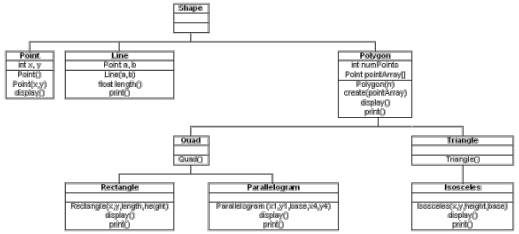
Reuse: 1. Automation

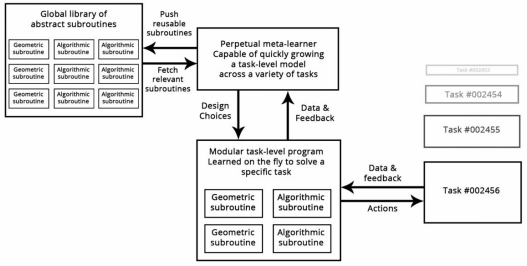




Software Design

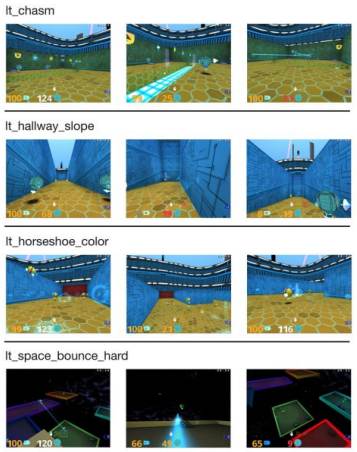
(in OOP)





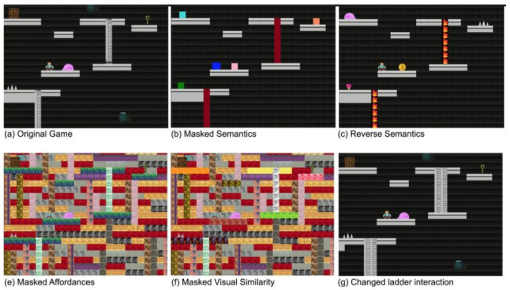
Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

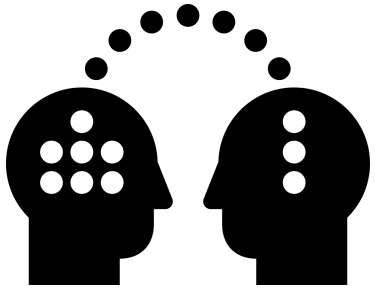
Reuse: 2. Automation

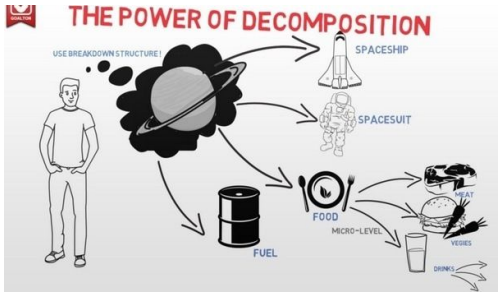


Transfer Learning

(simultaneous & sequential;
with & without explicit tasks)

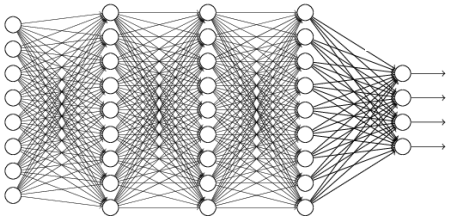


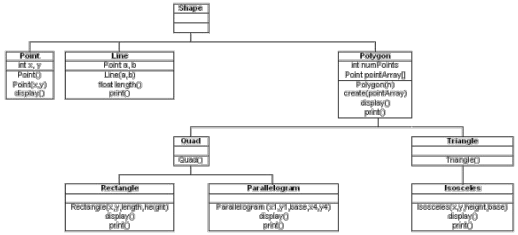




Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Both Frameworks





Strengths of Machine Learning:

- Highly Complex Rules
- Less Domain Knowledge
- Data-based
- Parameter Level Reusability & Fine-tuning

Strengths of Traditional Software:

- Module Level Reusability
- Tests / Debugging / Metrics
- Version Control
- Interpretability

Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

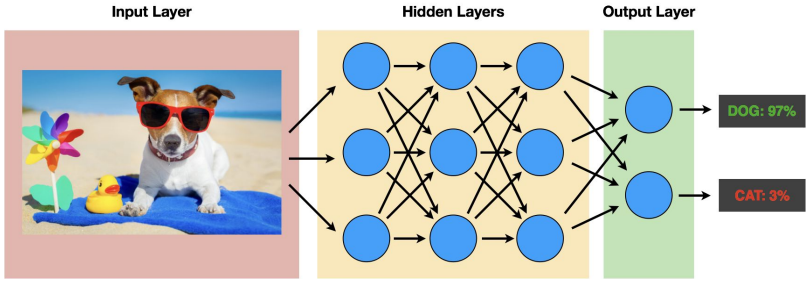
Strategies

- Module Level
  - **Consistent & Grounded Interfaces**
  - Decomposition
- Parameter Level
  - **Quicker Learning**
  - Slower Forgetting

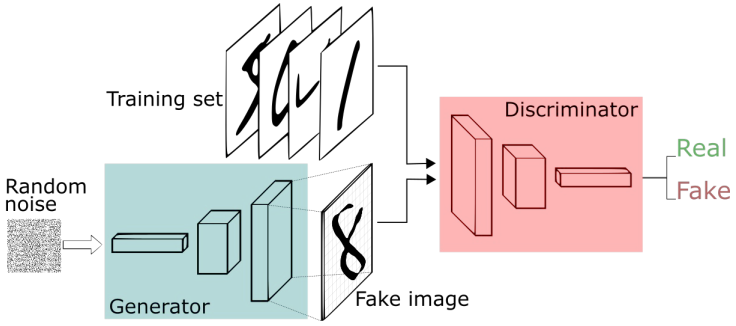
Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Model Interfaces

Multi Layer Perceptron



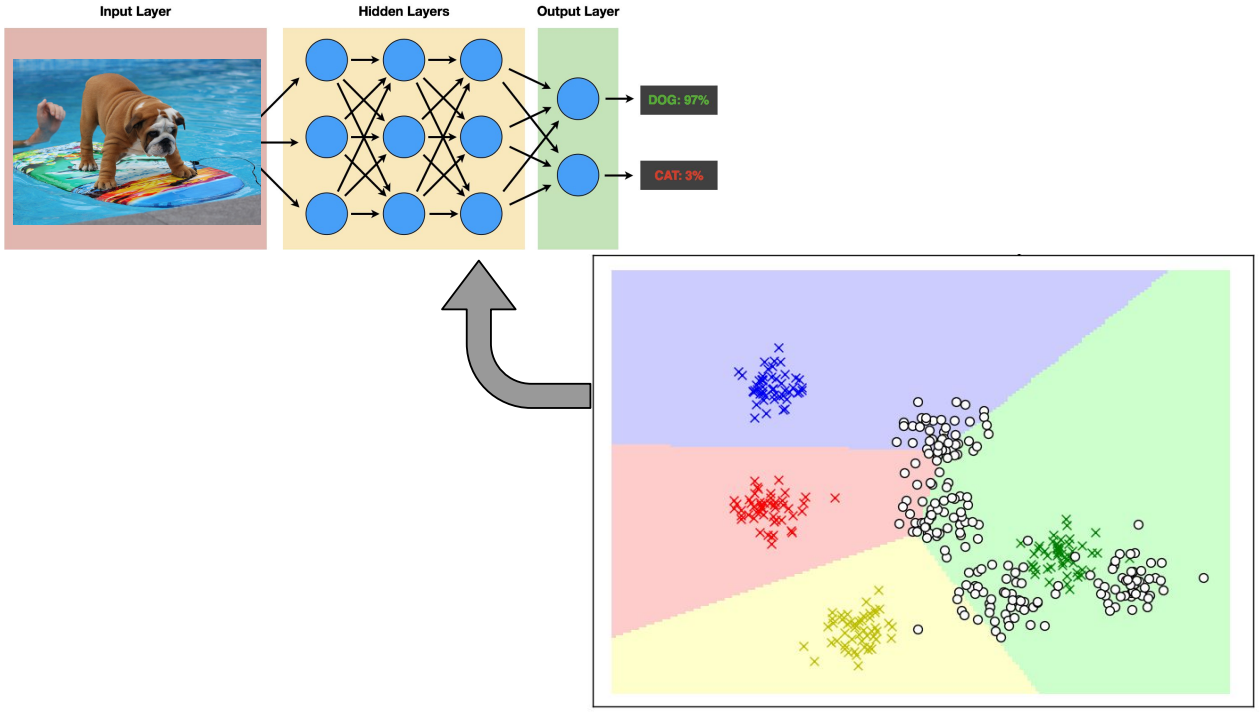
Generative Adversarial Network



Source: deeplearning4j

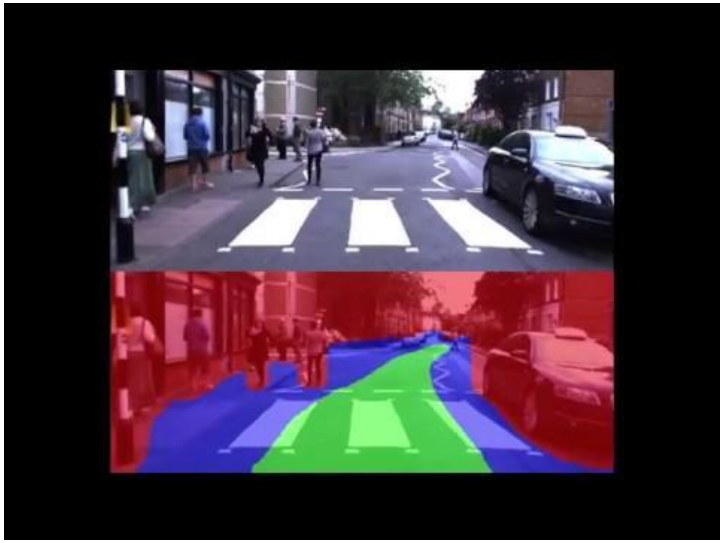
Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Consistent Interfaces



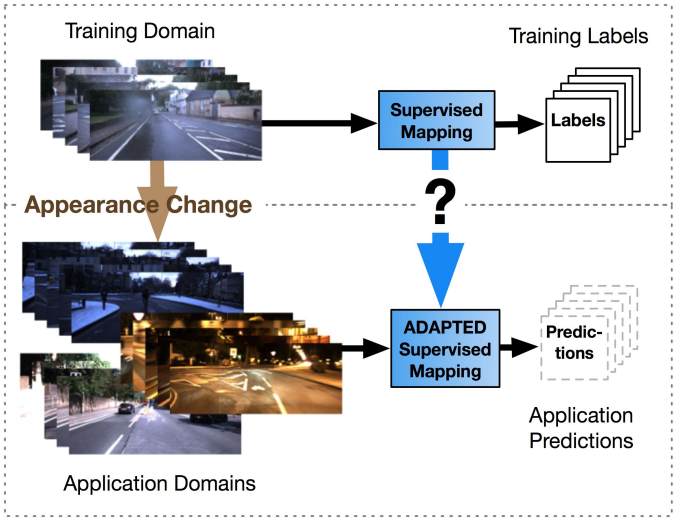
Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Domain Adaptation



Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

ML under Appearance Changes

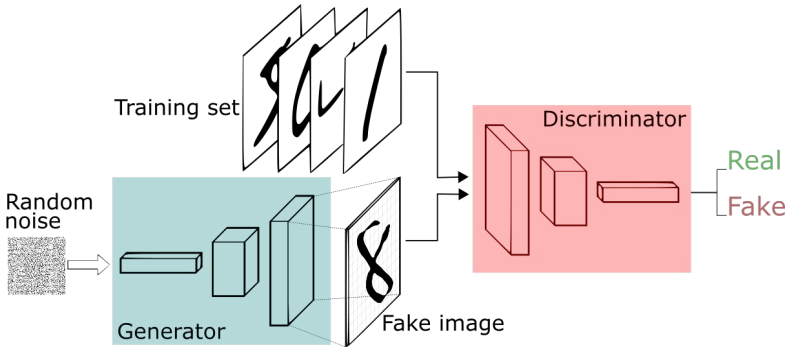


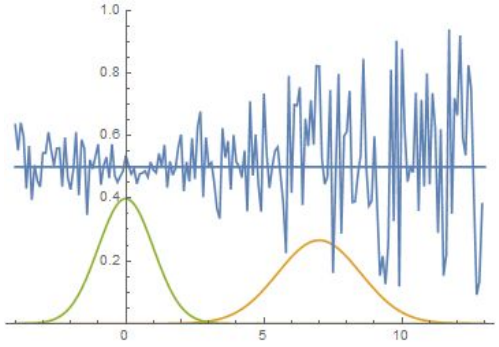
Optimisation of machine learning models to **perform well in the training domain**

However, **accuracy decreases under appearance changes** resulting in domain shifts.

Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Background: Adversarial Networks

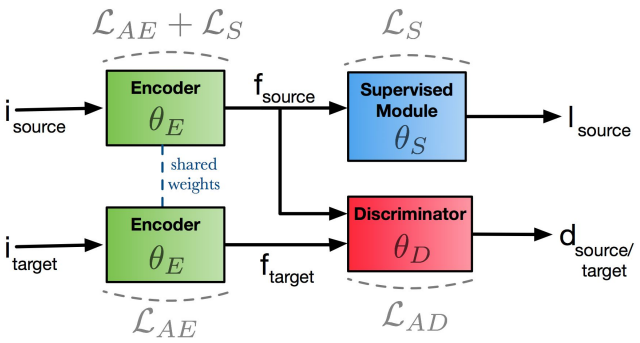




Source: deeplearning4j

Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Adversarial Domain Adaptation

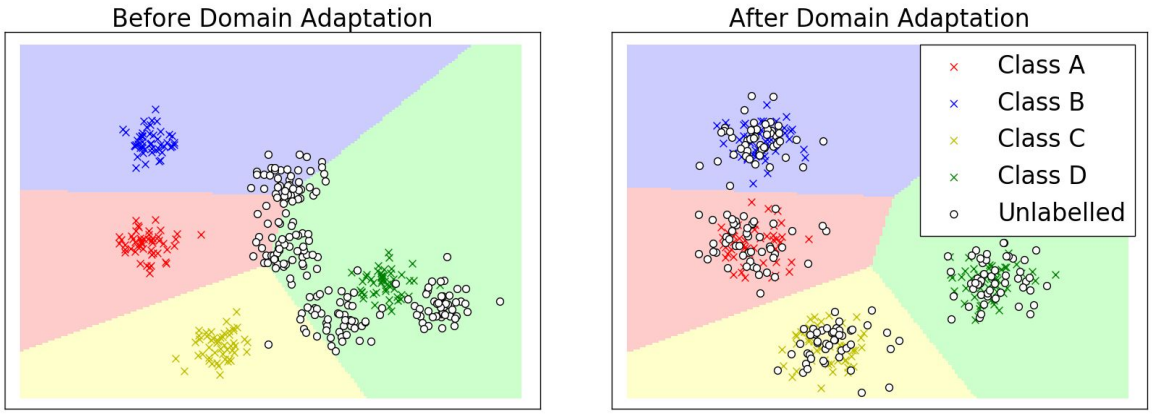


Optimise discriminator to determine the domain underlying samples in feature space

Adapt **encoder** to:

- **Minimise** discriminator performance
- **Maximise** feature distribution alignment

Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk



Aligning marginal distributions can align conditionals, given structural similarities between domains.

- The less difference in appearance - the easier this unsupervised alignment will we - curriculum for alignment

Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Evaluation

Scenarios

Day - Overcast

Day - Sunny

Night

Classification

	domains	AlexNet	AlexNet w ADA	AlexNet w target labels *
P_T[%]	overcast-sunny	67.95±1.02	**82.03**±2.09	(87.96±1.40)
P_T[%]	day-night	26.79±4.90	**30.21**±4.94	(90.42±1.04)

Free-Space Segmentation

	FCV-VGG16	FCN-VGG16 w ADA	FCN-VGG16 w target labels *
P_T[%]	75.12±0.76	**85.27**±1.03	(93.94±0.84)

ADA **outperforms** in the target domain in all test cases.

However, the approach gains its relevance for domains of **stronger visual similarity**.

Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Continuous Deployment

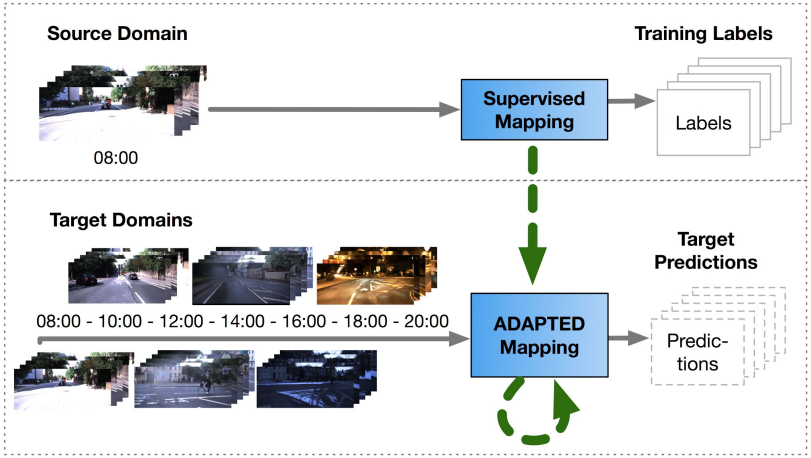
Day -
Sunny

Day -
Overcast

Night

Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

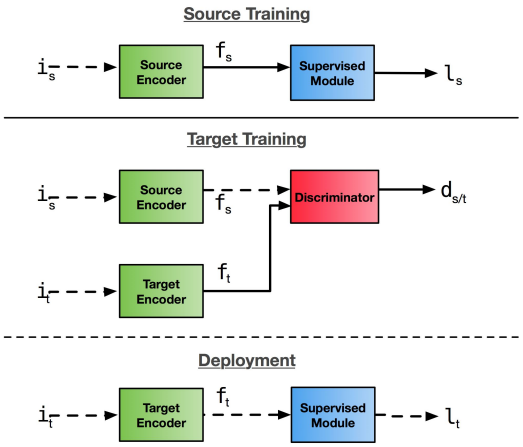
Incremental Domain Adaptation



Incrementally adapt over the minor appearance changes
which accumulate to significant domain shifts.

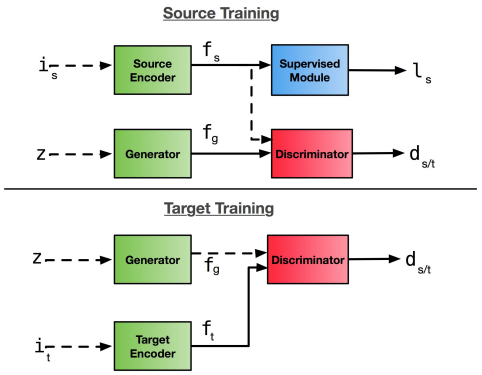
Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Regular



Dashed lines →
No back propagation

Simulated Source Domain



Discard the dependency on storing large amounts of annotated source data

Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Datasets

Incremental
Squeeze MNIST

Oxford RobotCar
Illumination Changes

Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Evaluation

target domains	only source	ADA	ADA SDM	IADA	IADA SDM
0.9	99.31	-	-	99.61	99.52
0.8	99.20	-	-	99.53	99.36
0.7	98.40	-	-	99.20	99.01
0.6	93.51	-	-	95.68	95.11
0.5	84.11	87.10	86.83	89.90	89.51

TABLE I: Target classifier accuracy on incrementally transformed MNIST dataset.

target domains	only source	ADA	ADA SDM	IADA	IADA SDM
morning	91.62	-	-	91.60	91.77
midday	90.70	-	-	91.05	90.50
afternoon	89.10	-	-	89.91	89.53
evening	87.08	-	-	89.01	87.34
night	76.27	78.67	77.12	80.21	79.37

TABLE III: Mean average precision results for segmentation task in continuous deployment scenario.

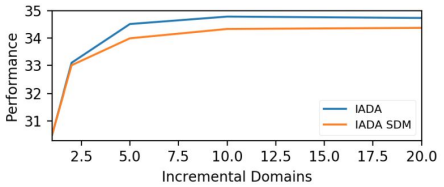


Fig. 5: Classifier accuracy of IADA in final target domain with varying number of intermediate domains for horizontal compression of 0.3.

IADA improves performance over regular one-step approaches

SDM significantly reduces memory requirements, but only reduces performance slightly in both cases

Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Take-Aways

- ML models only perform well for in-distribution samples
- Alignment of feature distributions improves performance
- Access to incremental shifts should be utilised

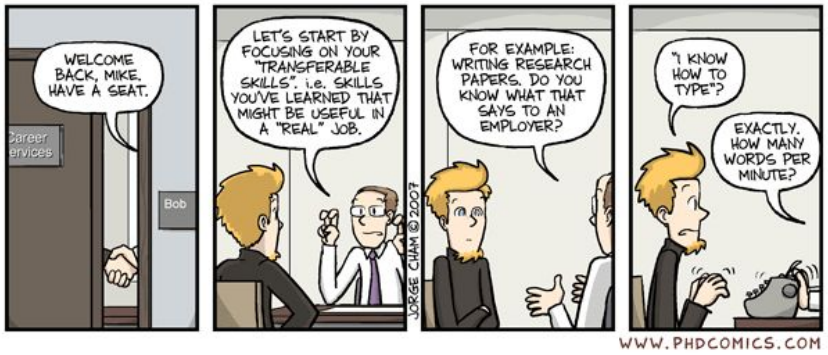
Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Strategies

- Module Level
  - Consistent & Grounded Interfaces
  - Decomposition
- Parameter Level
  - **Quicker Learning**
  - Slower Forgetting
- **Effect/Behaviour Level**

Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Simulation & Transfer Learning



Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

we want **real** robots

to perform **complex** tasks

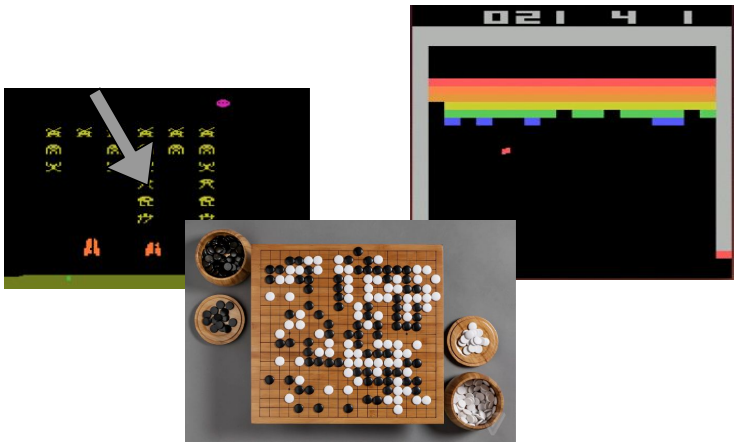
in the **real** world

Source: WIllow Garage

Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Solution: **Reinforcement Learning ?!**

Success !

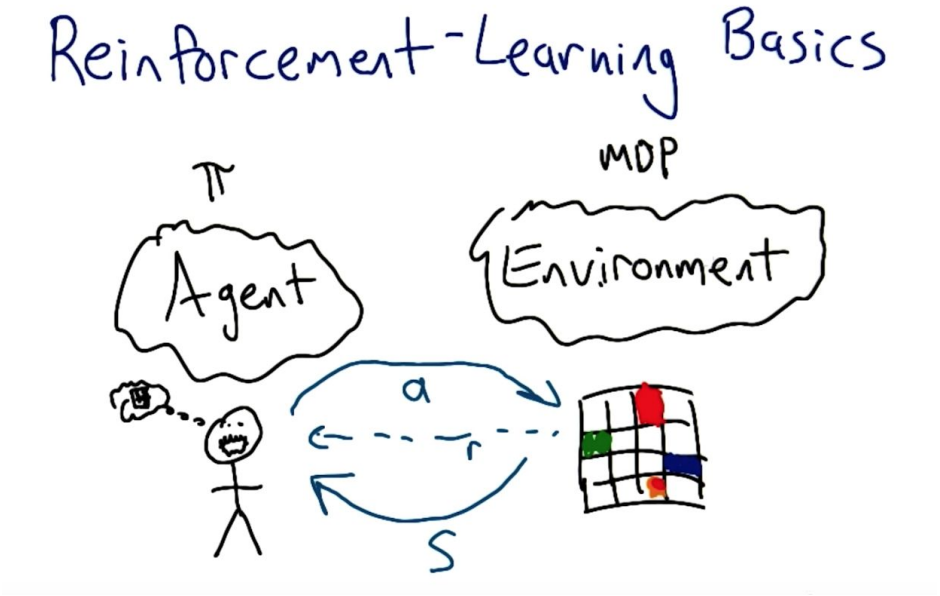


Source: DeepMind

… however all with access to **accurate simulation!**

Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Background: Reinforcement Learning



Sources: Udacity Reinforcement Learning

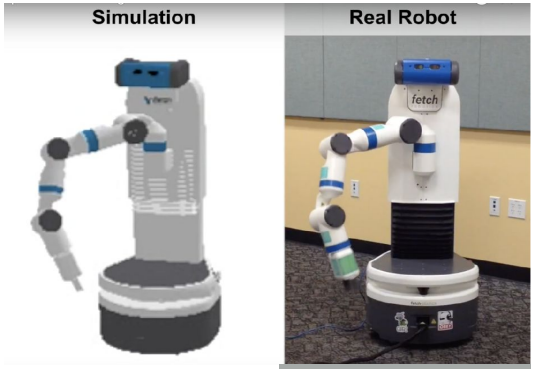
Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Motivation

Simulations have many advantages :

- fast
- safe
- cheap
- repeatable
- standardisable

However, while some models are useful, **all are wrong!**

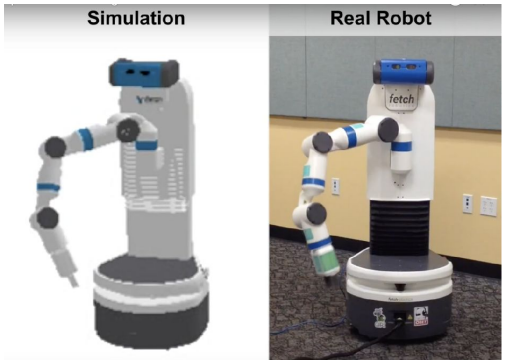


Sources: Sermanet et al 2016

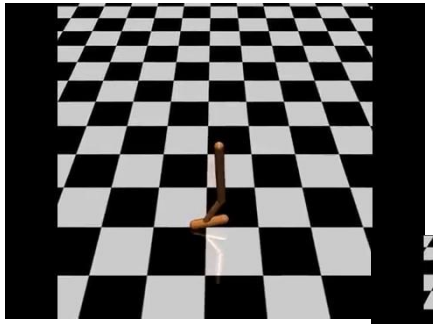
Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Challenges

Observation Space

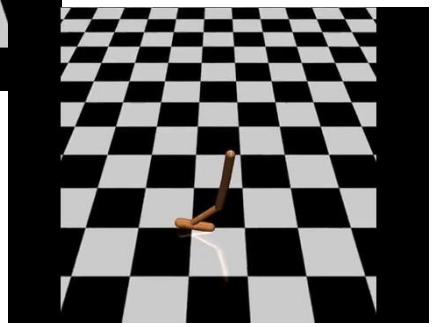


System Dynamics



Training Environment

Test Environment (higher mass)

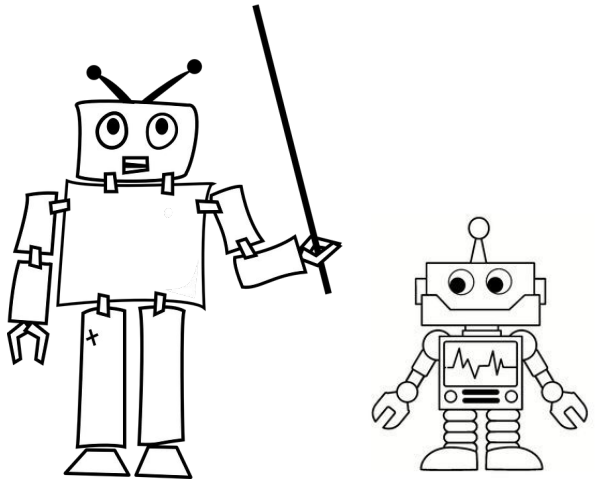


Other agents…

Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Behaviour Transfer

- Most typical form of transfer: **pretraining**
- The agent's parameters represent **only one type of reusable knowledge**
- Introduce new **type of transfer between agents in different domains**



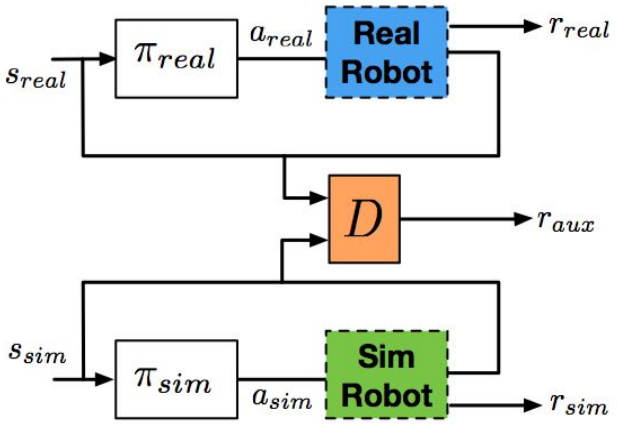
Source: pixabay.com

Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Mutual Alignment Transfer Learning

Method:

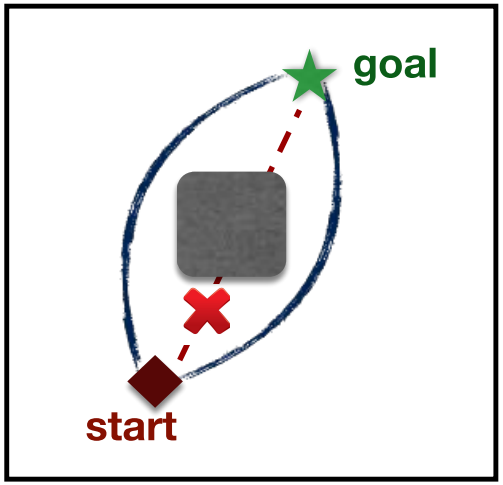
- **Parallel training** in simulation and real platform
- **Continuous alignment** of both agent's state distributions over visited states
- **Auxiliary rewards** to guide robot training via mutual alignment for both agents
- Straightforward to combine with other kinds of transfer (e.g. **fine-tuning**)



Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Multi-Modal Alignment Rewards

- Maximum likelihood objectives can lead to infeasible paths
- Density-based auxiliary objectives enable the representation of diverse & multimodal solutions



Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Experiments

Guiding Questions:

- **Benefits** with for the target system?
- Importance of **mutual vs unilateral alignment?**
- Capability to handle **more complex transfer scenarios**?

Methods:

independent real world training of π_{real}

direct transfer of π_{sim}

fine tuning of π_{sim} in the real world

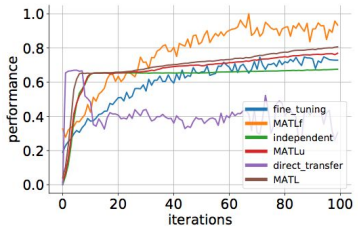
MATLu - unilateral alignment

MATL - mutual alignment

MATLf - combined with fine tuning

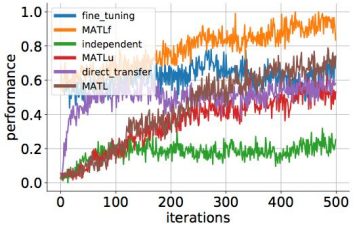
Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Sparse Rewards



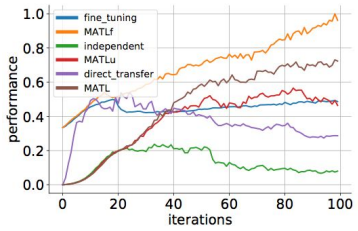
Cartpole Swingup

Reacher2D



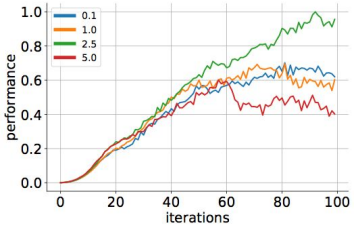
- improvements when training with only sparse rewards

Uninformative Rewards



Hopper2D

Hopper2D
- alignment weight

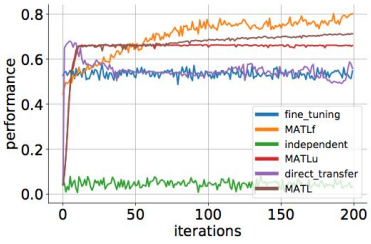


- improvements when training with uninformative rewards
- importance of alignment weight given conflicting rewards

details to all experiments can be found in the paper and online under https://sites.google.com/view/matl/

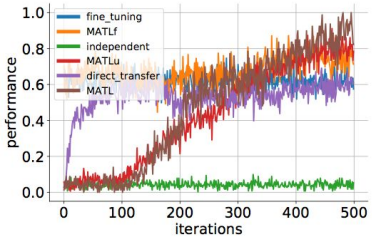
Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

No Environment Reward



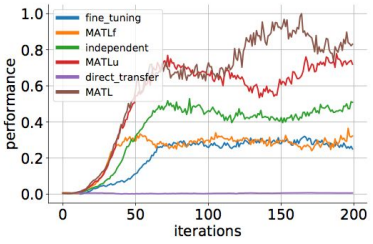
Cartpole Swingup

Reacher2D



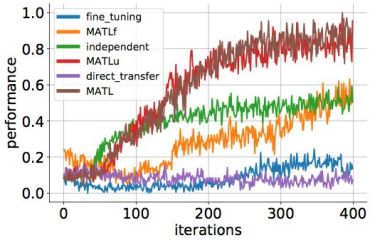
- improvements for training without environment rewards

MuJoCo to DART



Hopper2D

Reacher2D



- improvements even given procedural differences (when fine-tuning fails)
- MATLf demonstrates the same weakness as fine tuning

Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Take Aways

- Simulation-based teacher models can drastically accelerate training
- Alignment wrt. non-differentiable spaces possible via RL
- Bidirectional transfer outperforms unilateral

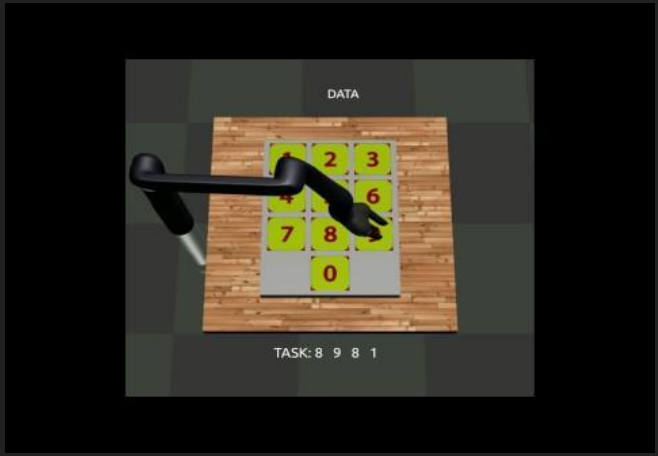
Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

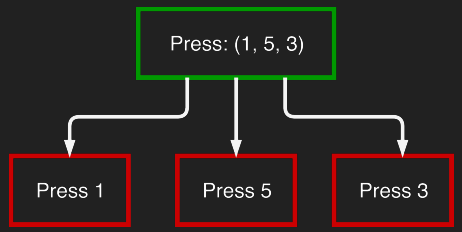
Strategies

- Module Level
  - Consistent & **Grounded Interfaces**
  - **Decomposition**
- Parameter Level
  - Quicker Learning
  - Slower Forgetting

Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Decomposition & Modular Learning





Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

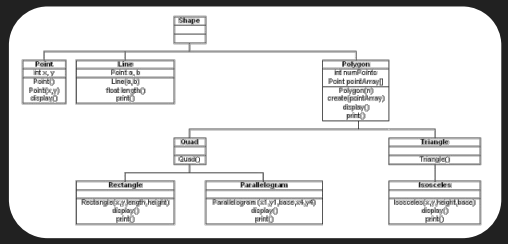
Modularity and Decomposition

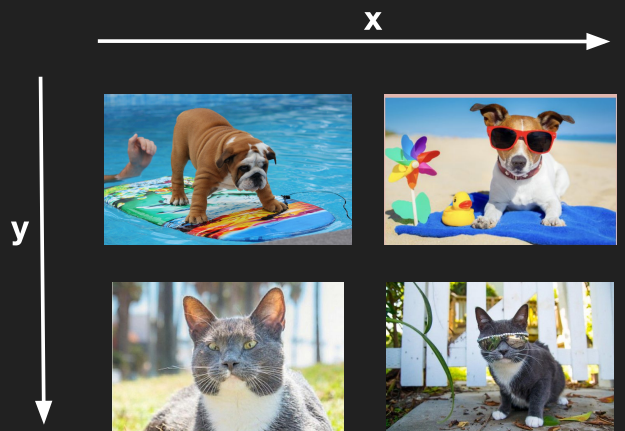
Advantages:

- Bidirectional Transfer
- Reusability
- Data efficiency
- Interpretability

Challenges:

- Learning to Decompose
- Learning to Generate





Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Learning from Demonstration (LfD)

In:

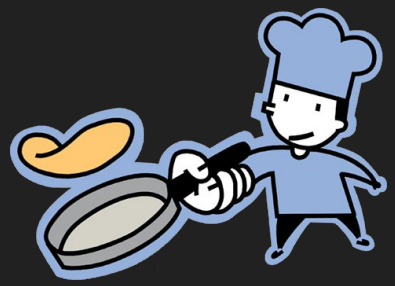
1. Demonstrations of length T

$$\rho = ((s_1, a_1), (s_2, a_2), \ldots, (s_T, a_T))$$

Out:

- Control Policy

$$\pi(a|s)$$



Sources: Willow Garage, signspecialist.com

Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

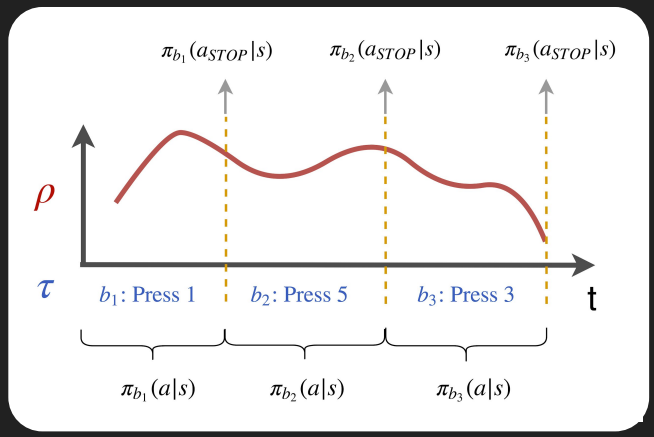
Modular LfD via Task Sketches

In:

1. Demonstrations of length T
   $\rho = ((s_1, a_1), (s_2, a_2), \ldots, (s_T, a_T))$
2. Task sketches of length L<<T
   $\tau = (b_1, b_2, \ldots, b_L)$

Out:

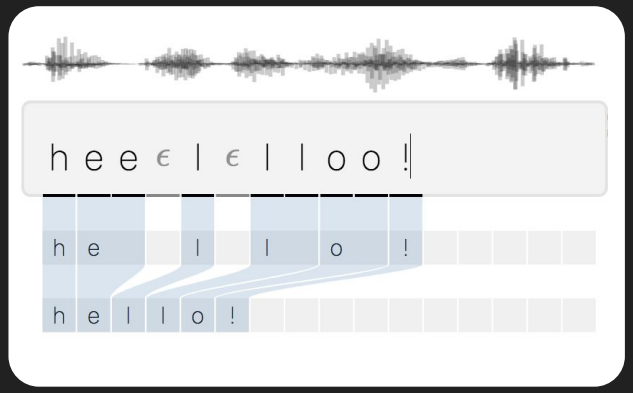
- Self-terminating sub-policies.
  $\pi_b(a^+|s) \quad a \in \mathcal{A}^+ = \mathcal{A} \cup a_{STOP}$

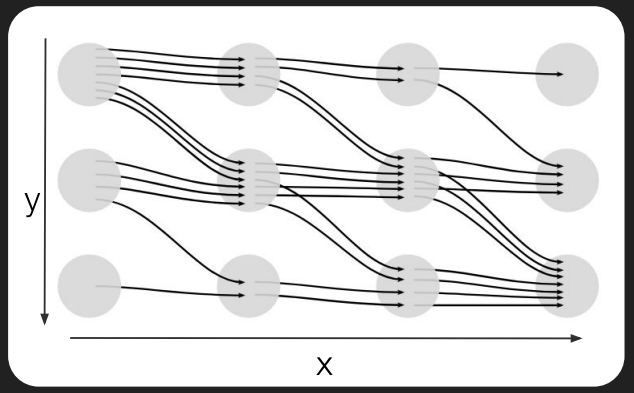


[1] Andreas, Jacob, Klein, Dan, and Levine, Sergey. *Modular multitask reinforcement learning with policy sketches*. In International Conference on Machine Learning, pp. 166–175, 2017.

Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Background: Connectionist Temporal Classification



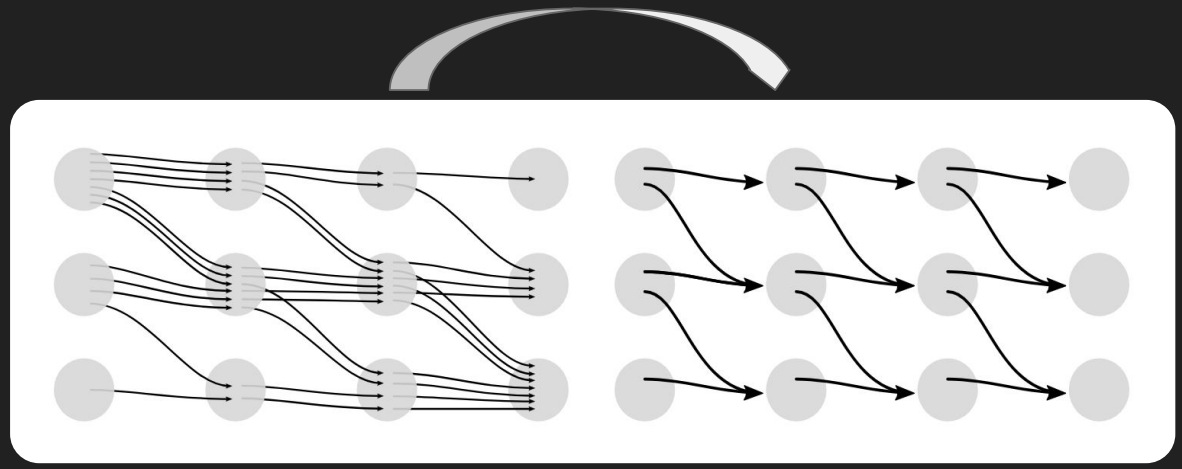


Sources: distil.pub

- Sequence to sequence model
- Applied in e.g. speech processing
- Per sequence max likelihood instead of per timestep

Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Background: Connectionist Temporal Classification



- Computation of all possible alignments is computationally expensive
- Dynamic programming simplifies the computation

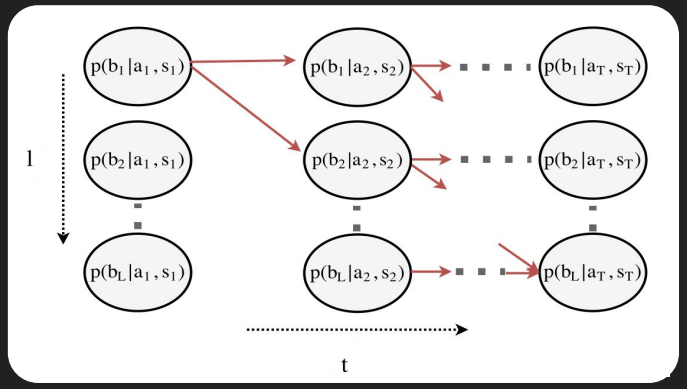
Sources: distil.pub

Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Naive Approach: Independent Alignment & Control

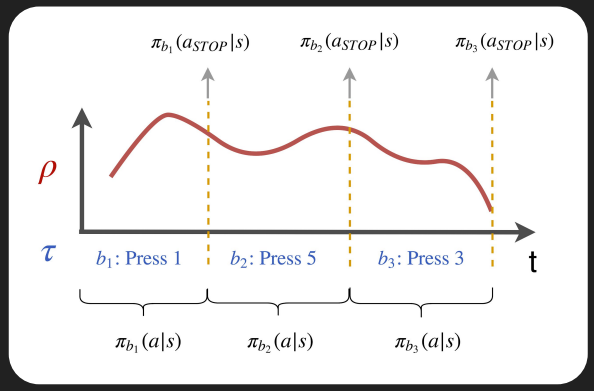
Step 1 - Alignment via CTC

$$\psi^* = \text{argmax}_\psi \, \mathbb{E}_{(\rho,\tau)} [p_\psi(\tau|\rho)]$$



Step 2 - Control via Behavioral Cloning

$$\theta^*_{k=1,\ldots,K} = \text{argmax}_{\theta_k} \, \mathbb{E}_{\rho_k} [\sum_{t=1}^{T_\rho} \log \pi_{\theta_k}(a_t^+|s_t)]$$



[1] Graves, Alex, Fernandez, Santiago, Gomez, Faustino, and Schmidhuber, Jurgen. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In International Conference on Machine Learning, pp. 369–376. ACM, 2006.

Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Temporal Alignment for Control (TACO)

Optimise Joint Likelihood : $\theta^* = \mathbf{argmax}_\theta \, \mathbb{E}_{\rho,\tau}[p(\tau, \mathbf{a}_\rho | \mathbf{s}_\rho)]$

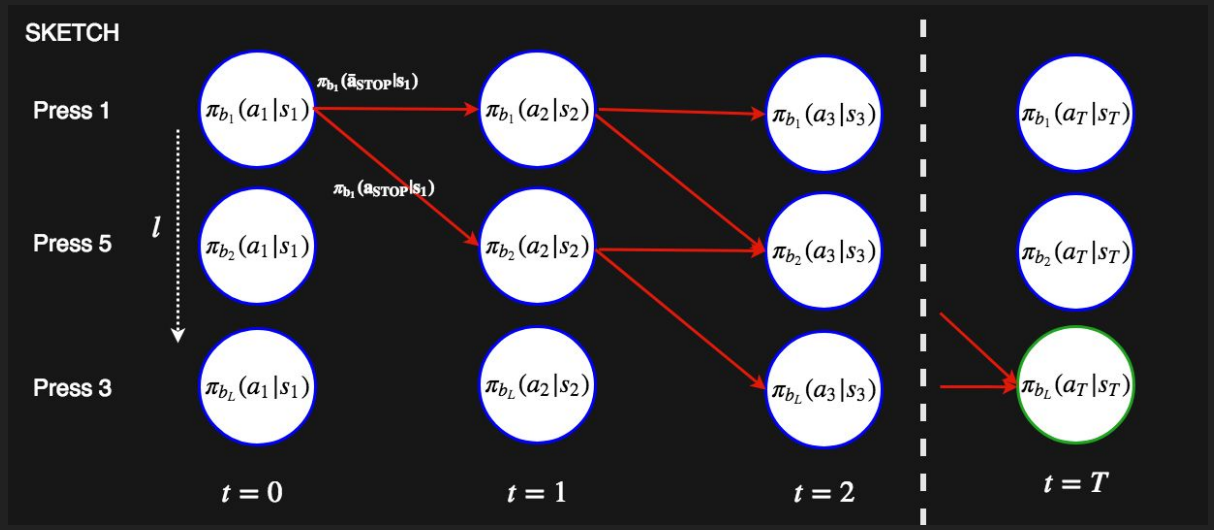
Advantages:

- Policy learning influences the alignment.
- All policies are exposed to all data points.
- Inductive bias for control: The objective reflects the fact that we need good policies as a result of the alignment.

Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

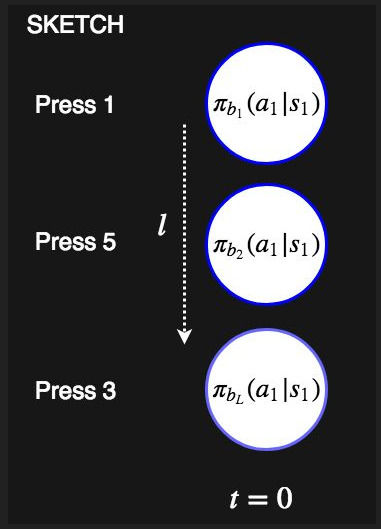
Temporal Alignment for Control (TACO)

Marginalise over all possible alignments:



Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Temporal Alignment for Control (TACO)

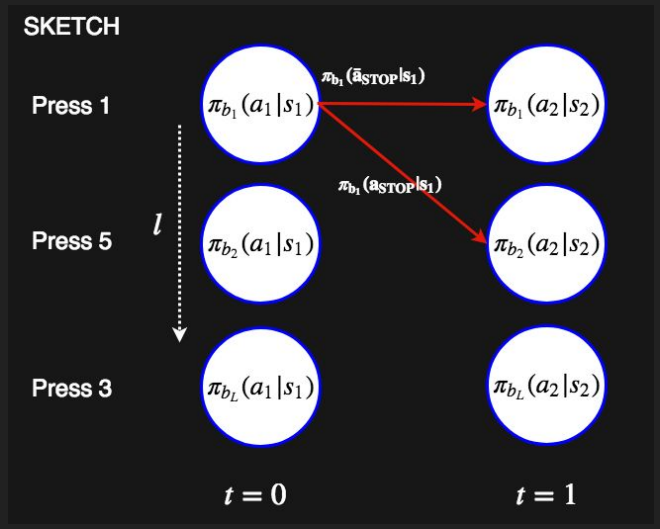


Forward variable initialisation:

$$\alpha_1(l) = \begin{cases} \pi_{\theta b_1}(a_1|s_1), & \text{if} \quad l = 1, \\ 0, & \text{o/w} \end{cases}$$

Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Temporal Alignment for Control (TACO)

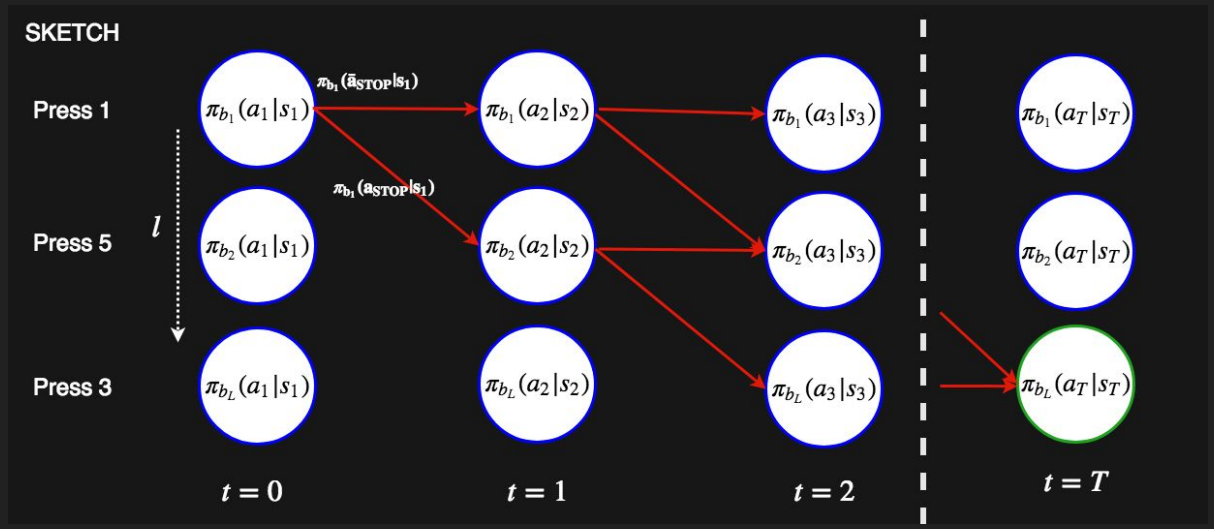


Recursion:

$$\alpha_t(l) = \pi_{\theta b_l}(a_t|s_t)\big[\alpha_{t-1}(l-1)\pi_{\theta b_{l-1}}(a_{STOP}|s_t) + \alpha_{t-1}(l)(1-\pi_{\theta b_l}(a_{STOP}|s_t))\big].$$

Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

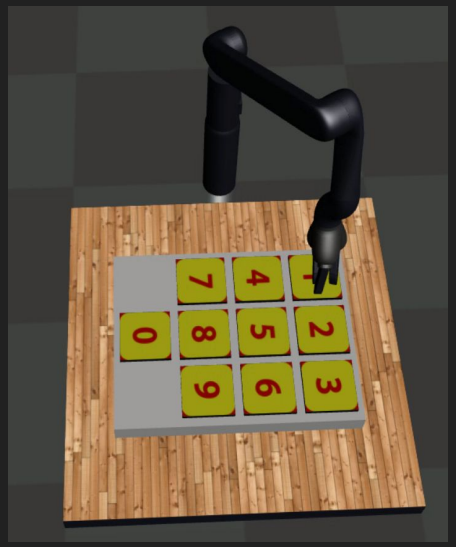
Temporal Alignment for Control (TACO)



$$\alpha_T(L) = p(\tau, \mathbf{a}_\rho | \mathbf{s}_\rho)$$

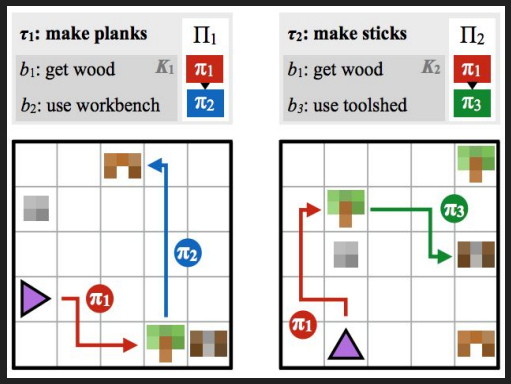
Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

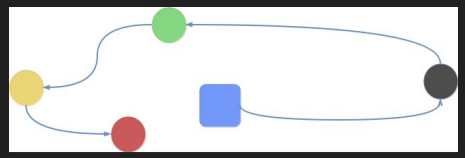
Evaluation



Jaco Dial

CraftWorld

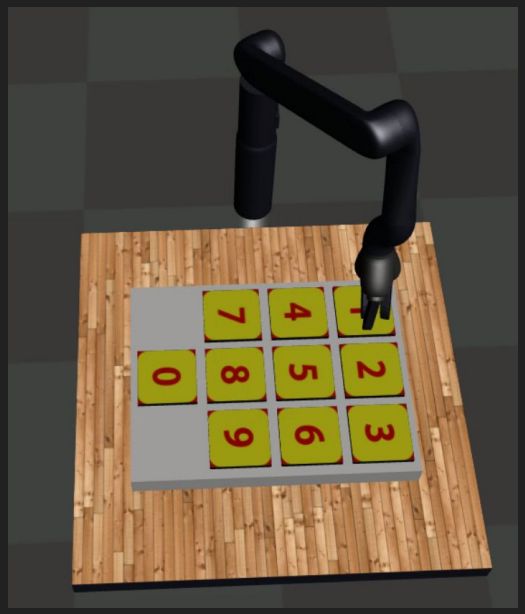




NavWorld

Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Dial Domain



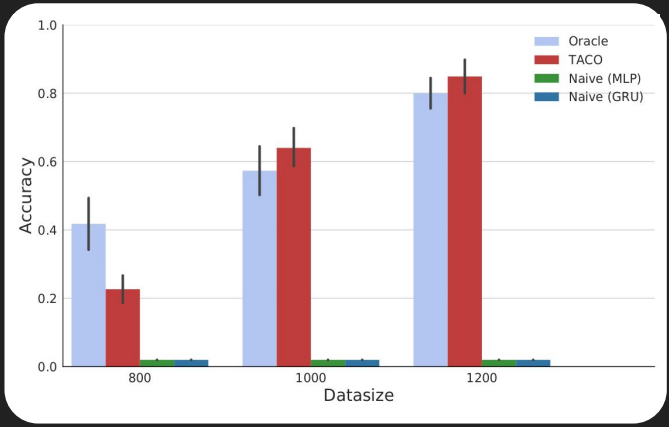
- 6 DoF Jaco Arm
- *Pin Code* demonstration trajectories
- Sketches: e.g. (0, 5, 6, 8)

State spaces:

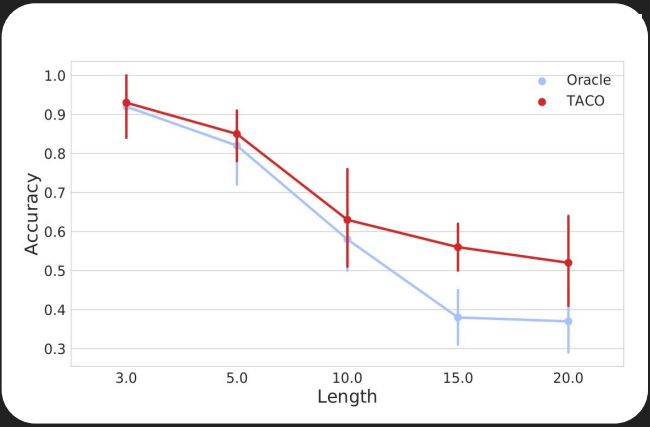
1. Joint angles + distances
2. **Only pixels**

Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Results: Dial Domain



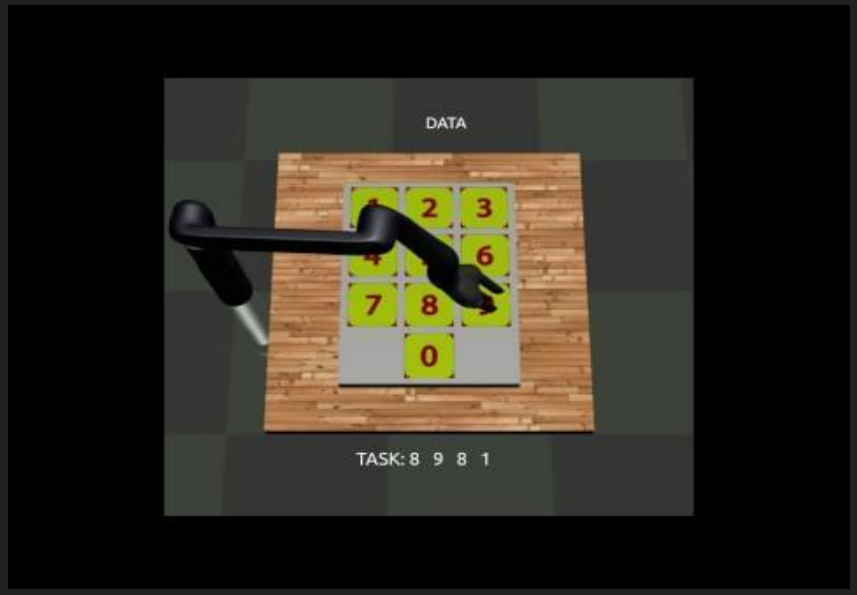
Increasing training data size



Increasing Test Sketch Sequence Length
(Fixed Train Sketch Length)

Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Dial Domain: Qualitative Results



Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Take Aways

- Weak instead of full supervision
- Decomposition enables transfer (& 0-shot)
- End-to-end optimisation for alignment and generation
- Intuitive interface for human operators via task dictionary

Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Reusable Learning

Goals:

- Efficient Learning
- Stronger Generalisation
- Forward and Backward Transfer

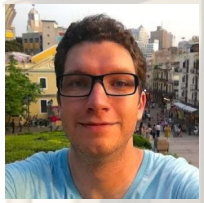
Challenges:

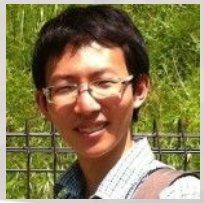
- Consistent Interfaces
- Task/Data Decomposition
- Catastrophic Forgetting
- Expressive Models/Flexible Learning
- Problem/Curriculum Design

Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Collaborators

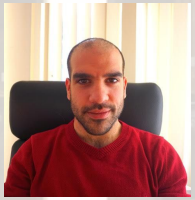






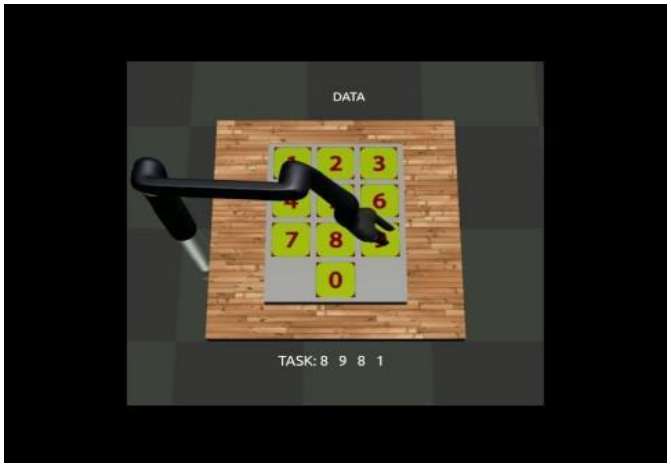





Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk

Thank you!



Markus Wulfmeier - University of Oxford - markus@robots.ox.ac.uk