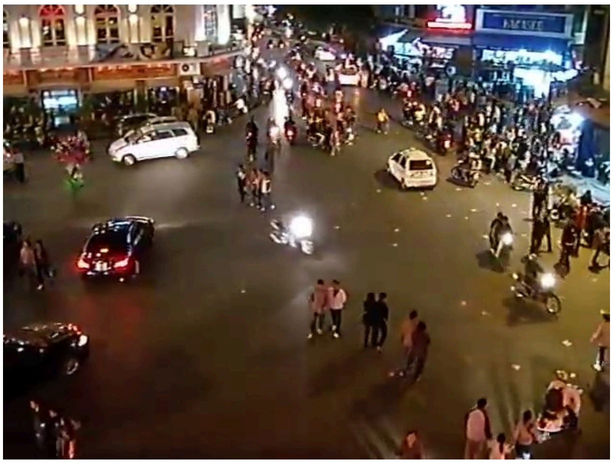
AI Transparency in Autonomous Vehicles

Justus Piater, University of Innsbruck

Table of Contents

1. Preface

1.1. Driving is a very complex activity.

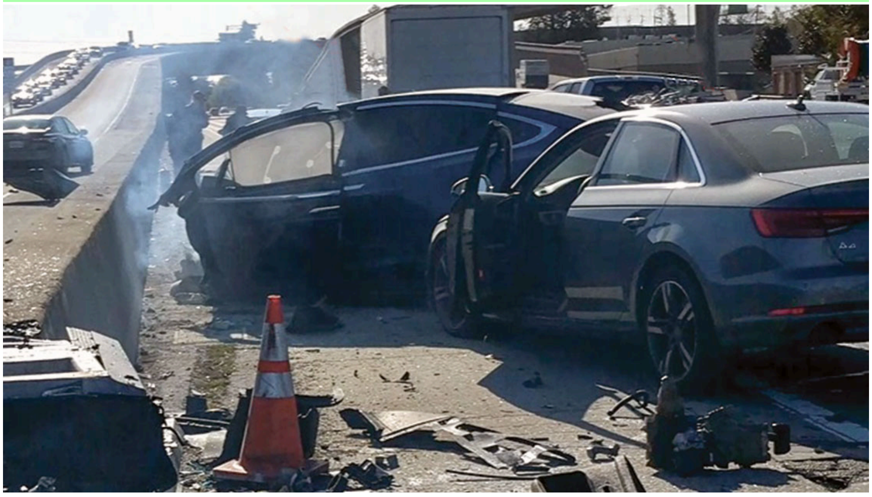


[mix75689 on Youtube[1]]

1.2. Driving is a high-stakes activity.

Like piloting aircraft, but distributed.

After failure we must find and fix the cause.



[Insurance Institute for Highway Safety, USA, 2018[2]]

"IIHS test drives of the Model S on public roads suggest Autopilot may be confused by lane markings and road seams where the highway splits."

[1] https://www.youtube.com/watch?v=NnUijTgk9rE
[2] https://www.iihs.org/news/detail/fatal-tesla-crash-highlights-risk-of-partial-automation

1.3. We want to be able to *trust* our automatic driver. (I)

If our car jolts sideways for no particular reason, we want to understand why. Otherwise we won't trust it.

[Josiah King on Youtube[3]]

1.4. We want to be able to *trust* our automatic driver. (II)

Having been driven impeccably through downtown rush hour and on highways, would I trust it to drive me … here?

Would I trust it to *refuse* to drive me here?

This is ***extrapolation***.

[RM Videos on Youtube[4]]

1.5. How can I trust my AI system?

- It worked yesterday ⇒ it will work today.
- If it *did not* perform impeccably in the past, or
  if I *cannot tell* whether it does the right thing:

  credit scoring

  - I want to understand the algorithm and how it was trained.

    This would often motivate me *not* to trust the system…

  - *I* want to understand what *the system* understands! (***AI Transparency***)
  - …

[3] https://www.youtube.com/watch?v=6y1e0skfJts
[4] https://www.youtube.com/watch?v=jPyYGw9Jn6w

1.6. Why don't I trust my AI system?

There is a ***mismatch*** between [Lipton 2016, Doshi-Velez and Kim 2017]

- the ***training objectives*** (prediction metrics)

and

- ***real-world cost*** (my life).

Real-world cost also depends on ***secondary factors*** (besides prediction metrics) that are often ***hard to model***:

- Causality (as opposed to just correlation)
  - E.g.: Use of *context* = both strength and weakness.

    Bushes to aid lane following
- Transferability
  - to unfamiliar situations (extrapolation outside the training set)
  - to adversarial environments
- Ethical considerations: fairness, …
- Informativeness
  - to humans, as decision support

1.7. How can I understand my AI system?

- Explain black-box models post-hoc
- Learn interpretable models

2. Explanation of black-box models

2.1. Explanation of Neural Networks for Image Analysis

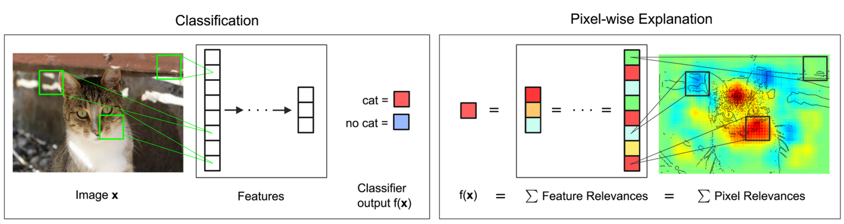
Heat map $h(x)$ of the *influence* of each pixel of input image x on output $g(x)$

- Gradient-based methods:

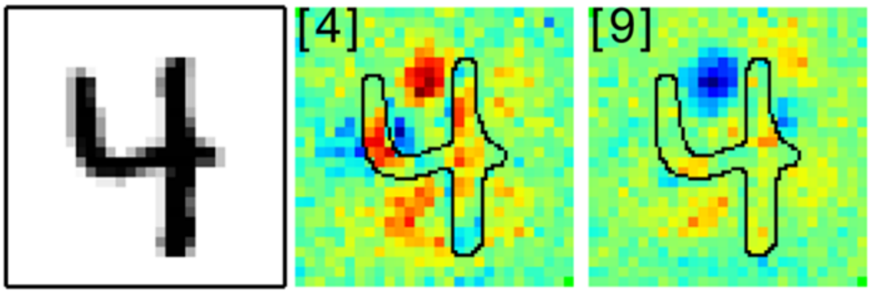
  $$h(x) = \nabla_x g(x)$$

- Layerwise Relevance Propagation (LRP) [Bach et al. 2015, Montavon et al. 2018]

  $$\begin{aligned} R_i^L &= g(x) \\ R_i^l &= \sum_j \frac{x_i^l (W^l)^+_{ji}}{\sum_k x_k^l (W^l)^+_{jk}} R_j^{l+1} \end{aligned}$$



2.2. LRP: Evidence Four and Against



[Bach et al. 2015]

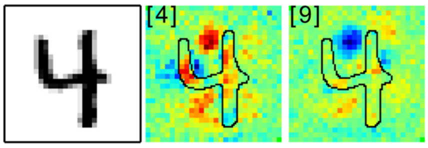
2.3. We know that classification can be attacked.

Single-pixel changes that affect the classification result [Su et al. 2019]

2.4. Explanations can be attacked too!

[Dombrowski et al. 2019]

2.5. Classical NN do not learn *Concepts*.



We parse digits into characteristic *strokes*.

We interpret *structure*, *function*, etc.

We would like the system to have a clear concept of *structure*, *support*, etc.

2.6. Classical NN have no idea what's going on.

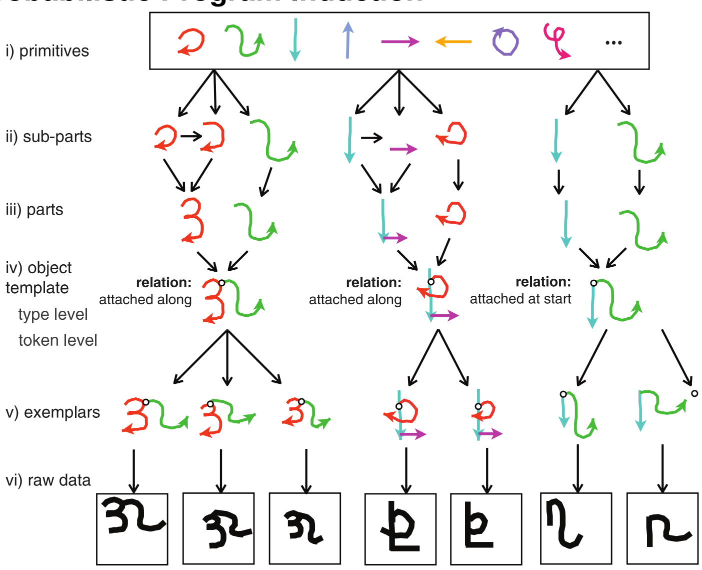
a woman riding a horse on a dirt road | an airplane is parked on the tarmac at an airport | a group of people standing on top of a beach

[Lake et al. 2017]

Image captions generated by a deep NN [Karpathy and Fei-Fei 2017; code[5]]

3. Interpretable Models

3.1. Probabilistic Program Induction



[Lake et al. 2015]

[5] https://github.com/karpathy/neuraltalk2

3.2. Probabilistic Program Induction

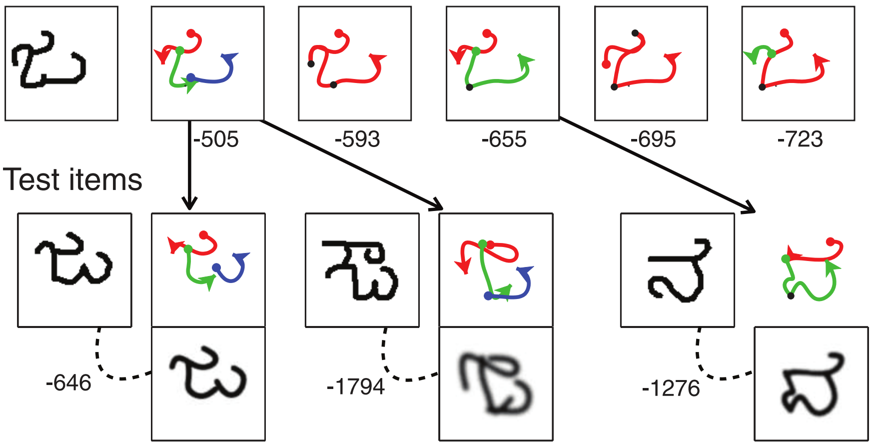
procedure GenerateType
  $\kappa \leftarrow P(\kappa)$ ▷ Sample number of parts
  for i = 1 ... κ **do**
    $n_i \leftarrow P(n_i|\kappa)$ ▷ Sample number of sub-parts
    for j = 1 ... n_i **do**
      $s_{ij} \leftarrow P(s_{ij}|s_{i(j-1)})$ ▷ Sample sub-part sequence
    end for
    $R_i \leftarrow P(R_i|S_1, ..., S_{i-1})$ ▷ Sample relation
  end for
  $\psi \leftarrow \{\kappa, R, S\}$
  return @GenerateToken(ψ) ▷ Return program

procedure GenerateToken(ψ)
  for $i = 1...\kappa$ **do**
    $S_i^{(m)} \leftarrow P(S_i^{(m)}|S_i)$ ▷ Add motor variance
    $L_i^{(m)} \leftarrow P(L_i^{(m)}|R_i, T_1^{(m)}, ..., T_{i-1}^{(m)})$
    ▷ Sample part's start location
    $T_i^{(m)} \leftarrow f(L_i^{(m)}, S_i^{(m)})$ ▷ Compose a part's trajectory
  end for
  $A^{(m)} \leftarrow P(A^{(m)})$ ▷ Sample affine transform
  $I^{(m)} \leftarrow P(I^{(m)}|T^{(m)}, A^{(m)})$ ▷ Sample image
  return $I^{(m)}$

[Lake et al. 2015]

3.3. Probabilistic Program Induction

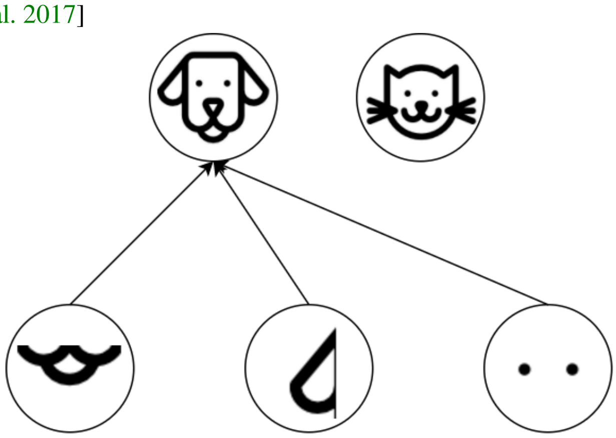
Training item with model's five best parses



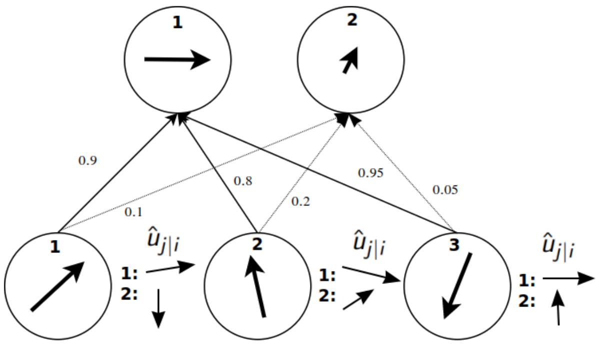
[Lake et al. 2015]

3.4. Capsule Networks

"A capsule is a group of neurons whose *activity vector* represents the instantiation parameters of a specific type of entity such as an object or an object part."
[Sabour et al. 2017]



3.5. Capsule Networks



Activity vectors $\mathbf{v}_j$

- Norm: probability that the entity exists
- Orientation: instantiation parameters of the entity
- By *squashing*, couplings c_{ij} are intended to form a ***parse tree***.

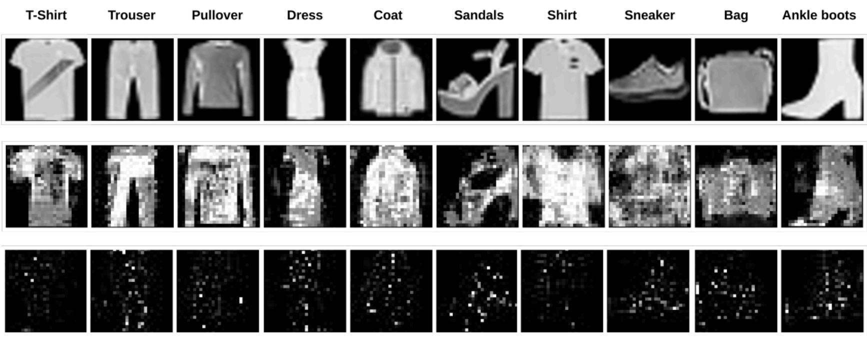
- Predictions $\hat{\mathbf{u}}_{j|i} = W_{ij}\mathbf{u}_i$ from lower capsule i to higher capsule j
- W_{ij} learned via backpropagation

Iterate (RBA):

$$
\begin{aligned}
c_{ij} &\leftarrow \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})} \\
\mathbf{s}_j &\leftarrow \sum_i c_{ij}\hat{\mathbf{u}}_{j|i} \\
\mathbf{v}_j &\leftarrow \frac{\|\mathbf{s}_j\|^2}{1+\|\mathbf{s}_j\|^2}\frac{\mathbf{s}_j}{\|\mathbf{s}_j\|} \\
b_{ij} &\leftarrow b_{ij} + \mathbf{v}_j^{\mathrm{T}}\hat{\mathbf{u}}_{j|i}
\end{aligned}
$$

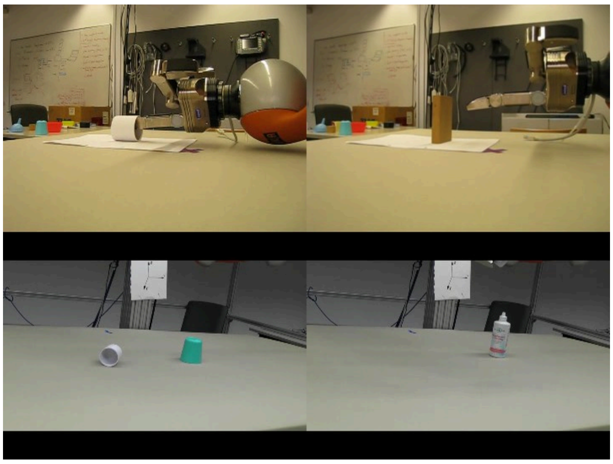
3.6. Sparse Parse Trees with γ-Capsules

- Original CapsNets do not produce sparse parse trees.
- γ-CapsNets do. [David Peer, Sebastian Stabinger, Antonio Rodríguez Sánchez; in progress]
  - Features are more human-interpretable.
  - Classification results are ***dramatically more robust to adversarial attacks*** than original CapsNets.



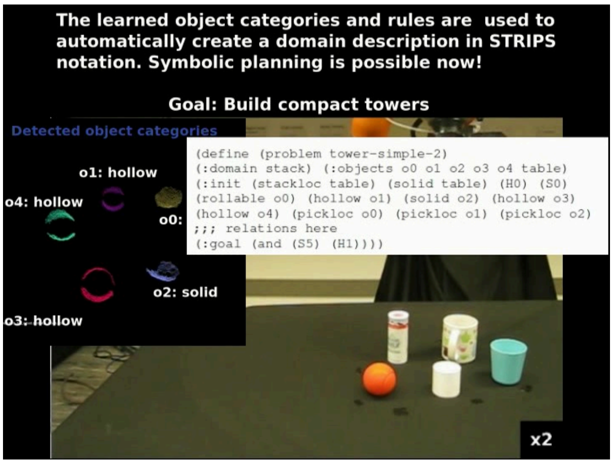
Top: Random training image. *Middle:* Average of 5 synthetic images optimizing that output capsule for γ-CapsNet. *Bottom:* Ditto for original CapsNet.

3.7. Learning Symbols From Sensorimotor Interaction



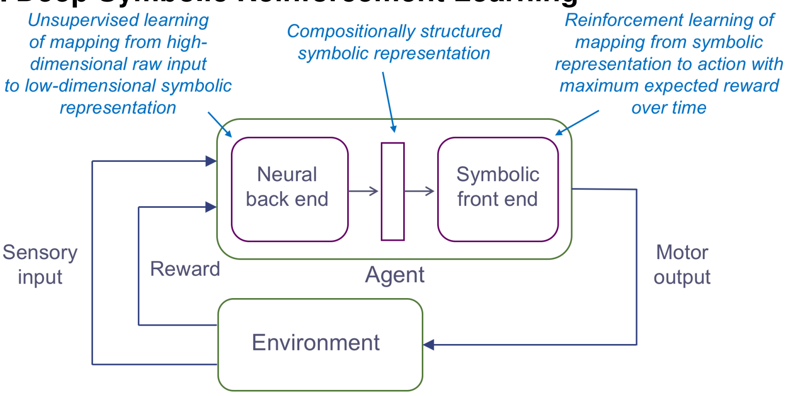
[Ugur et al. 2014]

3.8. Planning Using Learned Symbols



[Ugur and Piater 2015]

3.9. Deep Symbolic Reinforcement Learning

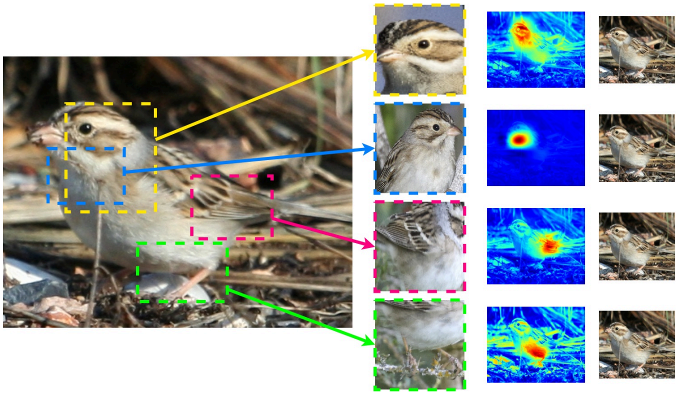


[Garnelo et al. 2016]

- ***Conceptual abstraction*** for transfer learning, planning, communication, …
- ***Compositional structure*** (here: probabilistic first-order logic)
- ***Common-sense priors*** and ***causal rules*** can be wired into the representation

3.10. Interpretable Models Can Be Powerful.

- Monolithic model with many parameters
  - powerful model without feature engineering
  - hard to interpret
- Structured model with many parameters [Rudin 2019]
  - powerful (but hard-to-interpret) backend learns interpretable concepts
  - interpretable (but powerful) frontend learns ultimate objective



4. Conclusions

4.1. What Can We Learn From (Human) Biology?

Computers are good at

- symbolic reasoning
- pattern classification and regression

Computers are poor at

- forming symbols
- functional understanding

Lessons:

- We should work on ***symbol formation / concept learning***.

  (Some approaches: clustering, autoencoders, CapsNets, Deep Symbolic RL)

- We should work on ***functional understanding***.

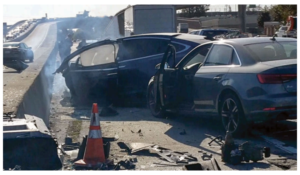
  (Some approaches: physics-based simulation; intuitive physics [Battaglia et al. 2013]; also builds on concepts)

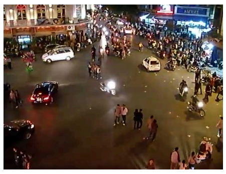
- The human visual system is not monolithic but is made up of ***specialized modules and pathways*** (dorsal/ventral, FFA, …)
  - Traffic signs should be OCR'ed.
- The human visual system is limited.
  - Success of autonomous vehicles hinges on ***sensors*** that outperform humans.

[Eykholt et al. 2018]

4.2. Conclusion

- Vision is more than ML on pixels.
- The "Vision Problem" cannot be solved without solving the "AI Problem".
- Unless AI systems gain substantially more (*structural*, *causal*, *functional*, *cultural*) ***understanding***, I will not trust them to drive me here:





- Learned ***conceptual abstractions*** can go a long way towards *extrapolation* and *explanation* capabilities, building *performance* and *trust*.

5. References

5.1. References

S. Bach, A. Binder, G. Montavon, F. Klauschen, K. Müller, W. Samek, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation[1]". *PLOS ONE* 10, 2015.

P. Battaglia, J. Hamrick, J. Tenenbaum, "Simulation as an engine of physical scene understanding[2]". *Proceedings of the National Academy of Sciences* 110(45), pp. 18327–18332, 2013.

A. Dombrowski, M. Alber, C. Anders, M. Ackermann, K. Müller, P. Kessel, *Explanations Can Be Manipulated and Geometry Is to Blame*[3], 2019.

F. Doshi-Velez, B. Kim, *Towards A Rigorous Science of Interpretable Machine Learning*[4], 2017.

K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, D. Song, "Robust Physical-World Attacks on Deep Learning Models". *International Conference on Computer Vision and Pattern Recognition*, 2018.

M. Garnelo, K. Arulkumaran, M. Shanahan, "Towards Deep Symbolic Reinforcement Learning[5]". *Deep Reinforcement Learning Workshop*, 2016.

A. Karpathy, L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions[6]". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, pp. 664–676, 2017.

B. Lake, R. Salakhutdinov, J. Tenenbaum, "Human-Level Concept Learning through Probabilistic Program Induction[7]". *Science* 350, pp. 1332–1338, 2015.

B. Lake, T. Ullman, J. Tenenbaum, S. Gershman, "Building Machines That Learn and Think like People[8]". *Behavioral and Brain Sciences* 40, 2017.

Z. Lipton, "The Mythos of Model Interpretability[9]". *ICML Workshop on Human Interpretability in Machine Learning*, 2016.

G. Montavon, W. Samek, K. Müller, "Methods for Interpreting and Understanding Deep Neural Networks[10]". *Digital Signal Processing* 73, pp. 1–15, 2018.

C. Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead[11]". *Nature Machine Intelligence* 1, pp. 206–215, 2019.

S. Sabour, N. Frosst, G. Hinton, "Dynamic Routing Between Capsules". *Advances in Neural Information Processing Systems 30*, pp. 3856–3866, 2017.

J. Su, D. Vargas, K. Sakurai, "One Pixel Attack for Fooling Deep Neural Networks[12]". *IEEE Transactions on Evolutionary Computation* 23, pp. 828–841, 2019.

E. Ugur, S. Szedmak, J. Piater, "Bootstrapping paired-object affordance learning with learned single-affordance features[13]". *The Fourth Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics*, pp. 476–481, 2014.

E. Ugur, J. Piater, "Bottom-Up Learning of Object Categories, Action Effects and Logical Rules: From Continuous Manipulative Exploration to Symbolic Planning[14]". *International Conference on Robotics and Automation*, pp. 2627–2633, 2015.

[1] http://dx.doi.org/10.1371/journal.pone.0130140

[2] http://dx.doi.org/10.1073/pnas.1306572110
[3] https://arxiv.org/abs/1906.07983
[4] https://arxiv.org/abs/1702.08608
[5] http://arxiv.org/abs/1609.05518
[6] http://dx.doi.org/10.1109/TPAMI.2016.2598339
[7] http://dx.doi.org/10.1126/science.aab3050
[8] http://dx.doi.org/10.1017/S0140525X16001837
[9] https://arxiv.org/abs/1606.03490
[10] http://dx.doi.org/10.1016/j.dsp.2017.10.011
[11] http://dx.doi.org/10.1038/s42256-019-0048-x
[12] http://dx.doi.org/10.1109/TEVC.2019.2890858
[13] https://iis.uibk.ac.at/public/papers/Ugur-2014-ICDLEPIROB-119.pdf
[14] https://iis.uibk.ac.at/public/papers/Ugur-2015-ICRA.pdf