The Curious Robot: Learning Visual Representations via Physical Interactions

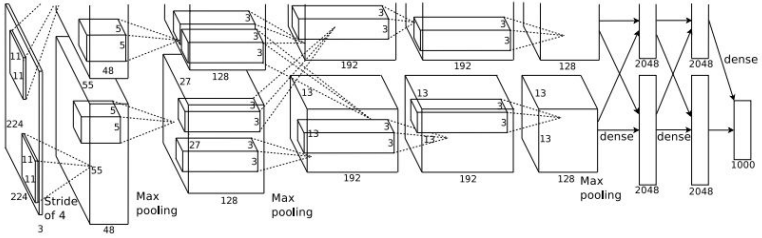
Presenter: Matthew Healy

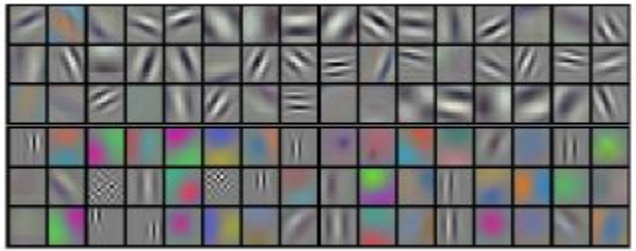
9/20/2022

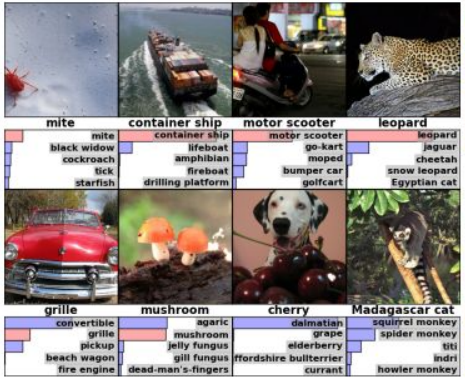
Motivation and Main Problem

What is the right supervisory signal to use?

Previous popular approaches: passive observations



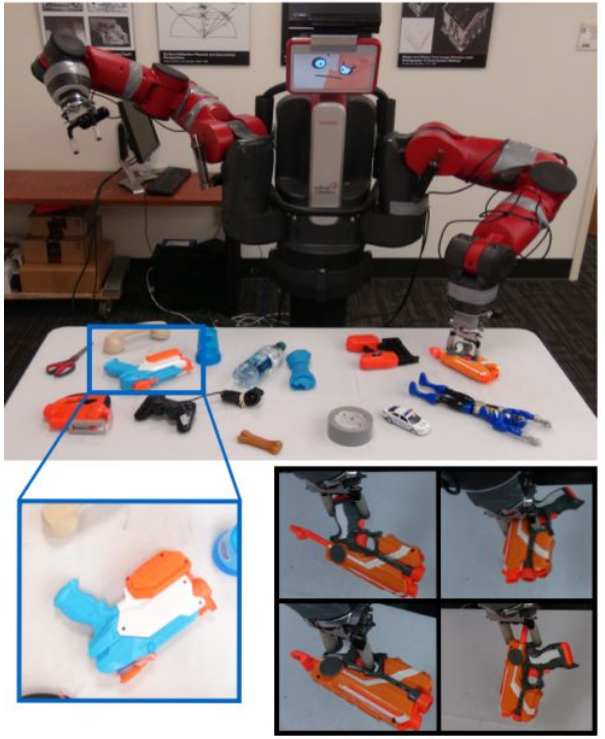




Motivation and Main Problem

What is the right supervisory signal to use?

Biological agents use physical interaction



Problem

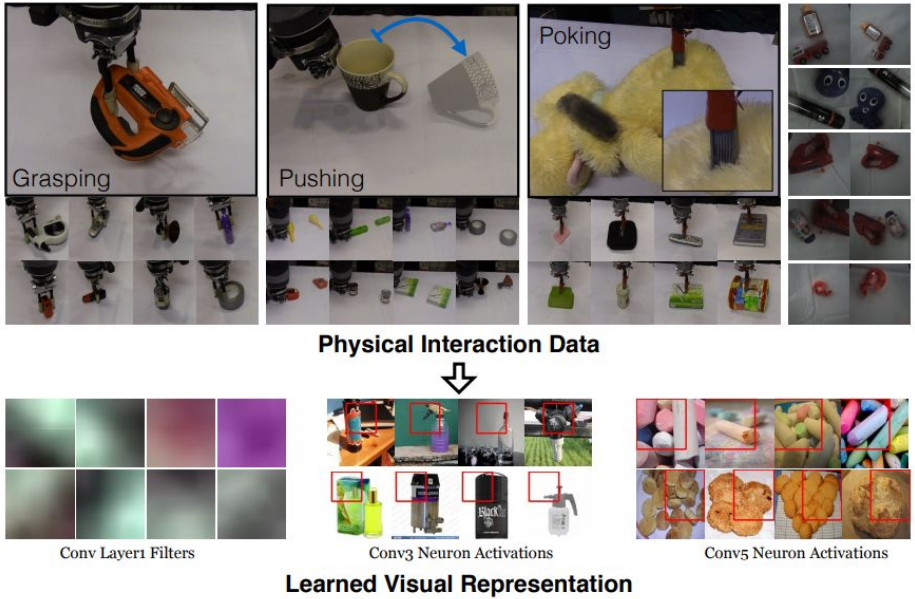
How do you learn a representation in an unsupervised manner and interact with the world for learning?

Provided physical interactions:

- Planar Grasps
- Planar Pushes
- Poking (Tactile Sensing)

Goal:

- Learn visual representations



Related Works - Active and Interactive Perception

- Active Vision. John Aloimonos et. al (1988)
  - Active observer can solve basic vision problems more efficiently than passive one
- Active Perception. Ruzena Bajcsy (1988)
  - Modeling and control strategies for perception
- Learning to See by Moving. Pulkit Agrawal et. al (2015)
  - Show benefit of egomotion for visual feature learning over class-label supervision
- Active Perception: Interactive Manipulation for Improving Object Detection. Quoc V. Le et. al (2010)
  - Method where robot moves in environment and manipulates object for detection

Prior Works - Unsupervised Learning

Both approaches only observe passive data

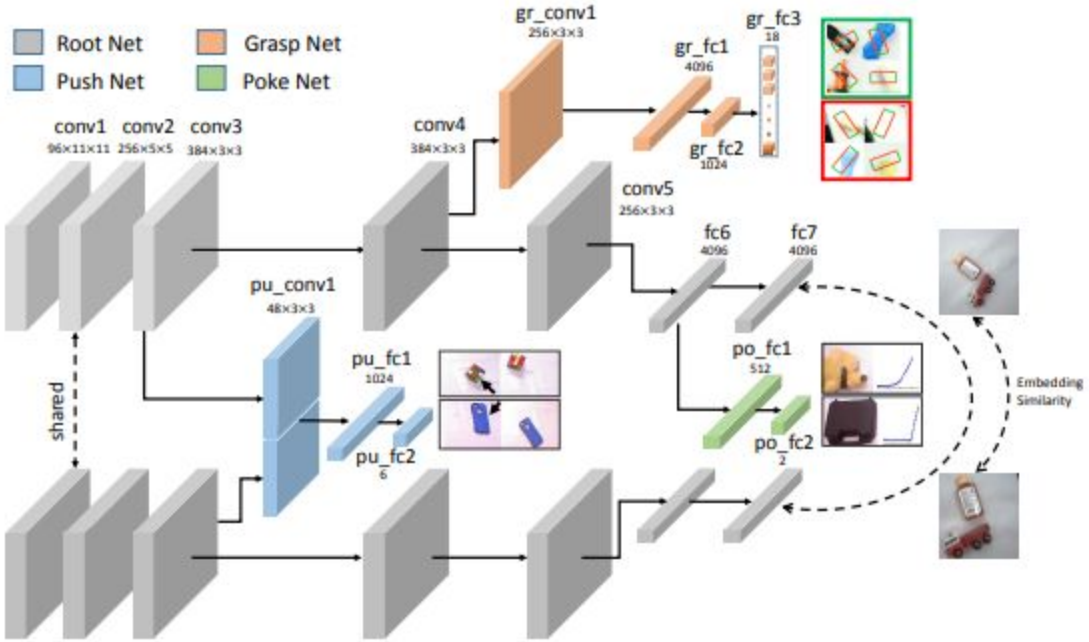
- Generative: Learning visual representations that can reconstruct images and are sparse. Recently used to generate realistic images, e.g. generative adversarial network (GAN) framework and its variants.
- Discriminative: Training a network on an auxiliary task where ground-truth is obtained automatically.

Example of GAN-Generated Photographs of Bedrooms.

Prior Works - Robotic Tasks

- **Grasping:** Often based upon complete knowledge of objects to be grasped, e.g. complete 3D model, surface friction, and mass distribution. Difficult to extract these attributes from RGBD cameras.
- **Pushing:** Aligning objects to reduce pose uncertainty before object manipulation. Relied on physics based models to simulate and predict required actions for desired change of object state.
- **Tactile Sensing:** Poking objects with skin sensor that measures pressure. Previously combined with computer vision for object detection.
- **Identity Vision:** Pairs of images in task's interaction contains images of objects with multiple viewpoints. Similar to idea of active vision where next best view chosen after inference.
- **Vision and Deep Learning for Robotics:** Using deep networks in robotic systems for grasp regression or learning policies for a variety of tasks.

Network Architecture



Grasp Network

Very similar to network in their earlier work

Input: Image of object

Output: 18D likelihood vector (18-way binary classifier)

Dataset: 43k grasp interactions from their earlier work

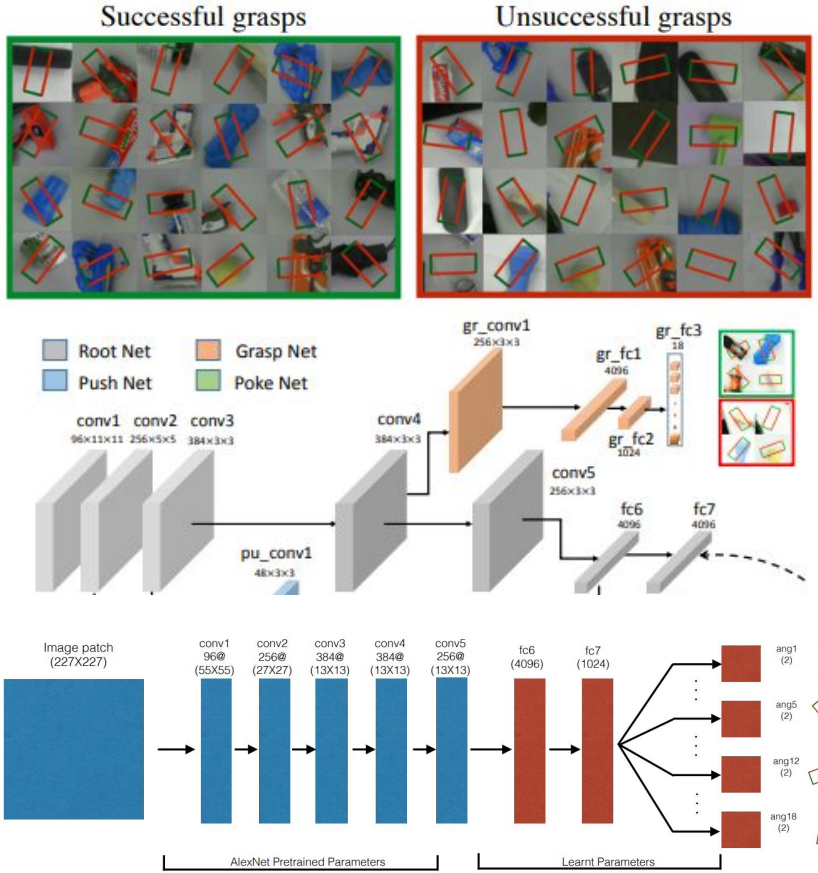
Network structure:

Root net (4 layers) + gr_conv1(256x3x3) + MP(3x3) + gr_fc1(4096) + gr_fc2(1024) + gr_fc3(18x2)

Training:

- RMSProp to back propagate before root net
- Gradients for root network stored and wait for aggregation
- Classification loss:

$$L = \sum_{i=1}^{B} \sum_{j=1}^{N=18} \delta(j, \theta_i) \cdot \text{softmax}(A_{ji}, l_i)$$



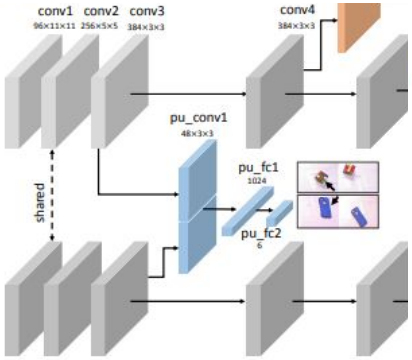
Push Network

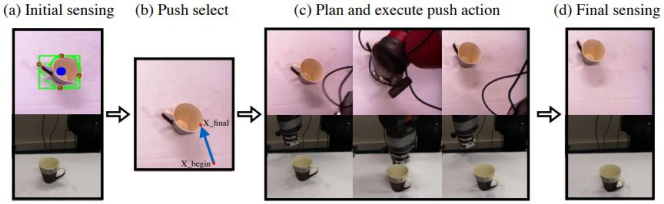
***Planar* Pushing** - no movement in-z axis

Input: Two images of object before and after pushing

Output: push-action $\{X_{begin}, X_{final}\}$

Dataset: 5k push actions on 70 objects using a Baxter robot



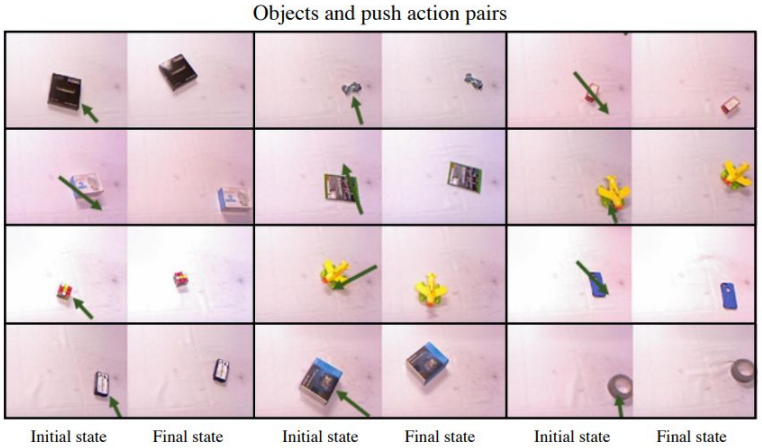


Network structure:

Siamese root net (3 layers) + pu_conv1(48x3x3) + pu_fc1(1024) + pu_fc2(5)

Training:

- RMSProp to back propagate before root net
- Gradient in pu_conv1 are accumulated and mean-aggregated before an update
- Gradients for root network stored and wait for aggregation
- Regression loss: mean squared error (MSE)



Poke Network

Measure pressure with tactile skin-sensor via voltage drop $\mathbf{p}_{do}$

Input: Image of object

Output: intercept and slope of tactile sensor plot described by $\mathcal{P}(\mathbf{p}_{do})$

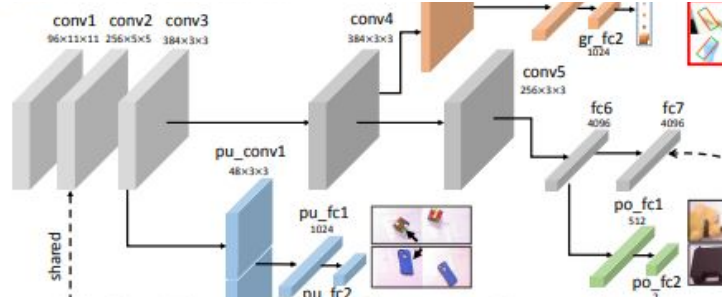
Dataset: 1k poke actions on 100 diverse objects using Baxter robot

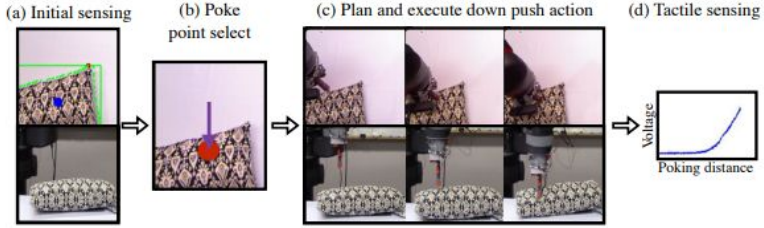
Network structure:

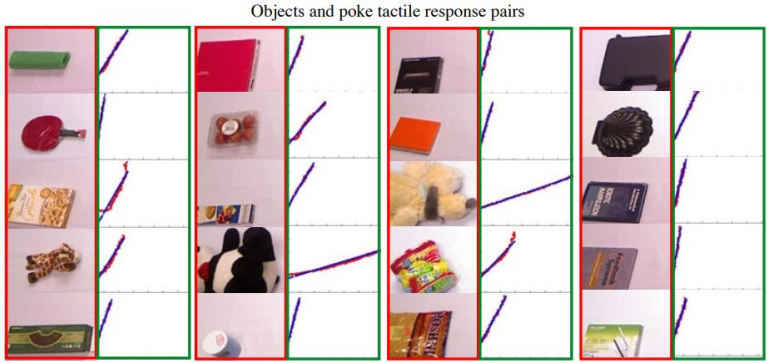
Root net (4 layers) + conv5(256x3x3) + MP(3x3) + po_fc1(512) + po_fc2(2)

Training:

- RMSProp to back propagate before root net
- Gradients for root network stored and wait for aggregation
- Regression loss: mean squared error (MSE)







Identity Similarity Embedding

Images of objects in the same task interaction should be closer in distance in fc7 feature space.

Input: Pair of images of same object

Output: fc7 feature representations

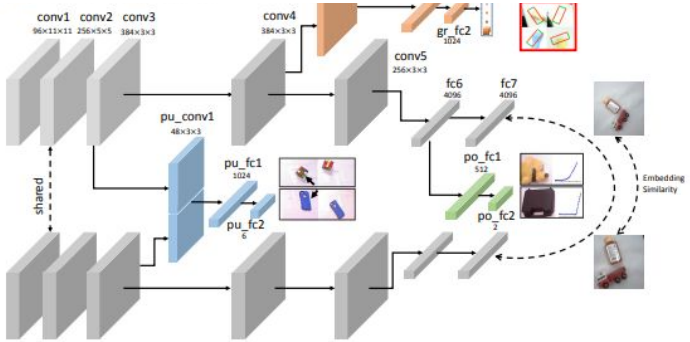
Dataset: 42k positive pairs of images and 42k negative pairs (images from different interactions)

Network structure:

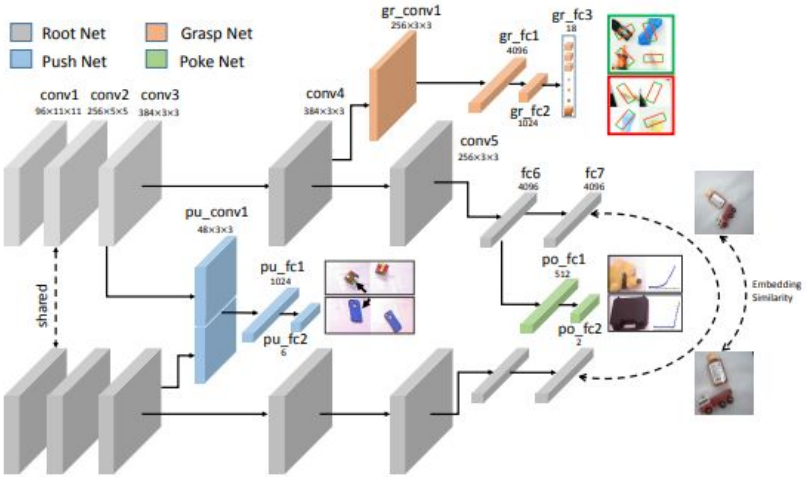
Root net (4 layers) + conv5(256x3x3) + MP(3x3) + fc6(4096) + fc7(4096

Training:

- Cosine embedding loss backpropagated through chain
- Gradients for two copies are accumulated and mean aggregated



Shared Network Architecture



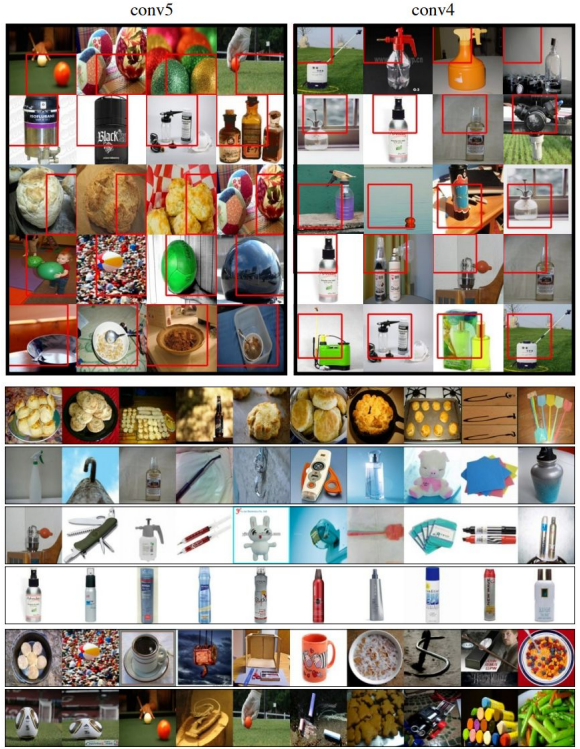
Root Network: common representation

Back propagation: Weights of root layers are aggregated and updated simultaneously

Training:

- Initialize root network and grasp network with Gaussian initialization
- Grasp network trained alone for 20k iterations
- Full architecture created with first conv4 copied from grasp learning
- Weights for subnets updated during respective backward propagation cycles while gradients for root net are accumulated and weight update step taken after each cycle of 4 task batches

Results



Experiment 1:

- 2500 Household ImageNet images
- Find images that maximally activate neurons
- conv5 able to correlate strong shape attributes

Experiment 2:

- 25 query images, 2500 as dataset
- conv5 feature space to perform nearest neighbors
- Nearest neighbors based on shape attributes

Classification Results

- 2500 Household ImageNet images
  - 100 each of 25 different objects
- UW RGBD dataset
- Caltech-256 dataset

- Correlation between robot tasks and semantic classification tasks
- Outperforms other unsupervised methods

Table 1. Classification accuracy on ImageNet Household, UW RGBD and Caltech-256

	Household	UW RGBD	Caltech-256
Root network with random init.	0.250	0.468	0.242
Root network trained on robot tasks (**ours**)	0.354	0.693	0.317
AlexNet trained on ImageNet	0.625	0.820	0.656
Root network trained on identity data	0.315	0.660	0.252
Auto-encoder trained on all robot data	0.296	0.657	0.280

Image Retrieval Results

- UW RGDB dataset
- fc7 features as visual representation
- Recall@k metric

Table 2. Image Retrieval with Recall@k metric

	Instance level				Category level			
	k=1	k=5	k=10	k=20	k=1	k=5	k=10	k=20
Random Network	0.062	0.219	0.331	0.475	0.150	0.466	0.652	0.800
Our Network	0.720	0.831	0.875	0.909	0.833	0.918	0.946	0.966
AlexNet	0.686	0.857	0.903	0.941	0.854	0.953	0.969	0.982

Task Ablation Results

- Trained network excluding 1 out of 4 tasks
- Suggests grasp task may be most important contribution to classification

Table 3. Task ablation analysis on classification tasks

	Household	UW RGB-D	Caltech-256
All robot tasks	0.354	0.693	0.317
Except Grasp	0.309	0.632	0.263
Except Push	0.356	0.710	0.279
Except Poke	0.342	0.684	0.289
Except Identity	0.324	0.711	0.297

Critique & Limitations

- Trained on planar tasks on tabletop, not easy to generalize to other settings
- Time consuming to gather physical interactions for individual objects
- No robust color information recognition, only shape information extracted
- Unclear if all tasks generally useful as input (e.g. push), task ablation only excludes 1 task at a time
- Difficult to tell how gradients from different tasks may interfere with each other since all tasks share same root network
- Biased towards grasping because network first trains only the grasp network and lower root network?

Table 3. Task ablation analysis on classification tasks

	Household	UW RGB-D	Caltech-256
All robot tasks	0.354	0.693	0.317
Except Grasp	0.309	0.632	0.263
Except Push	0.356	0.710	0.279
Except Poke	0.342	0.684	0.289
Except Identity	0.324	0.711	0.297

Future Works and Extended Readings

- Combining self-supervised tasks:
  - “Multi-Task Self-Supervised Visual Learning.” Carl Doersch et. al., 2017
  - “Cross-Domain Self-Supervised Multi-Task Feature Learning Using Synthetic Imagery.” Zhongzheng Ren et. al., 2018
- Learning through interaction:
  - “Learning to Poke by Poking: Experiential Learning of Intuitive Physics.” Pulkit Agrawal et. al., 2016
  - “Interactive Perception: Leveraging Action in Perception and Perception in Action.” Jeanette Bohg et. al., 2017
  - “Learning to push by grasping: Using multiple tasks for effective learning.” Lerrel Pinto et. al., 2017
  - “Grasp2Vec: Learning Object Representations from Self-Supervised Grasping.” Eric Jang et. al., 2018
  - “ViTac: Feature Sharing Between Vision and Tactile Sensing for Cloth Texture Recognition.” Shan Luo et. al., 2018
  - “Learning to Singulate Objects Using a Push Proposal Network.” Andreas Eitel et. al., 2019

Summary

- **Problem:** How do you learn a representation in an unsupervised manner and interact with the world for learning?
  - Presents method for learning visual representation from interactive physical tasks
  - Uses shared root network for 4 different tasks
- **Key Insights:**
  - Successfully combined robotic interaction and vision representation in manner opposite what was done previously
  - Results show correlation between robot tasks and semantic classification tasks
  - Grasping may be most important task for classification