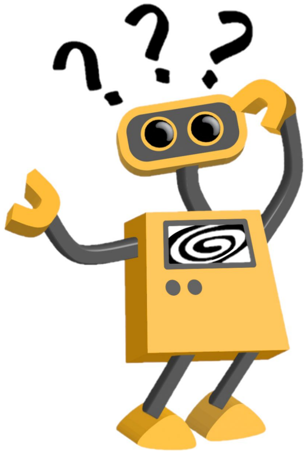
Exploration

EECS 598-12: Unsupervised Visual Learning
Presenter - Justin Bi, 4/12/2021

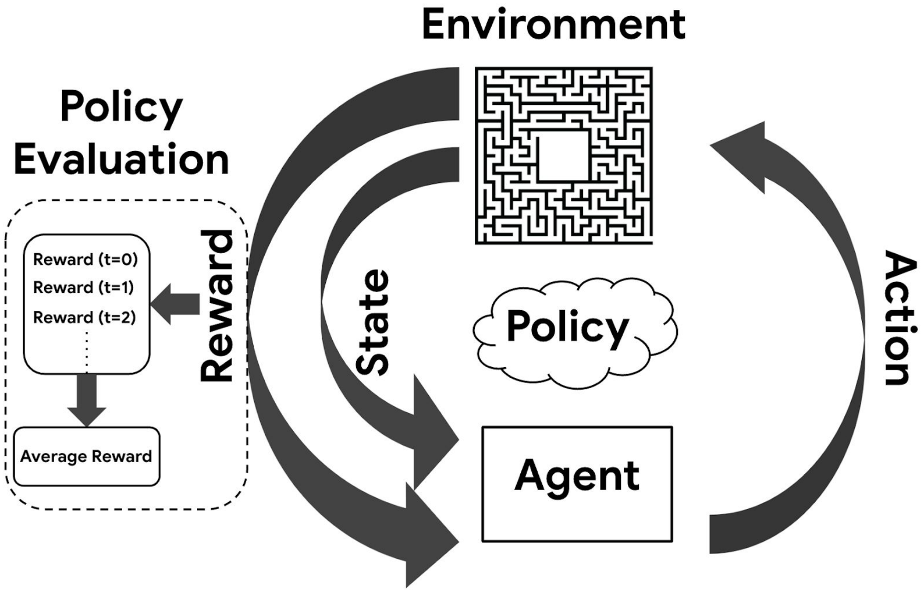
What is exploration, and why do we care?

- Exploration is the process of an agent learning about the environment it is operating in
- Greater knowledge leads to better-informed decision making in future tasks
- Unfortunately, difficult to solve with reinforcement learning



Reinforcement learning (RL) - quick refresher

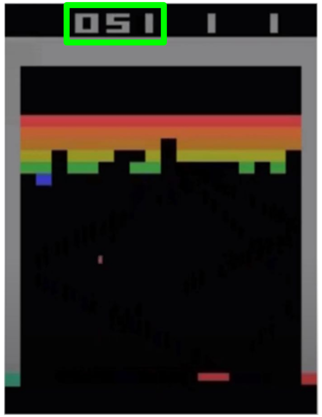
- Train an agent to interact with its environment in order to maximize rewards



https://ai.googleblog.com/2020/04/off-policy-estimation-for-infinite.html

Problem setup

- Using only *intrinsic* rewards, maximize the exploration performance of the agent by an *extrinsic* metric
  - Real world rewards are typically sparse or nonexistent



Game score is an example of extrinsic reward

Victoria Dean, Shubham Tulsiani, Abhinav Gupta, "See, Hear, Explore: Curiosity via Audio-Visual Association" NeurIPS 2020

See, Hear, Explore: Curiosity via Audio-Visual Association

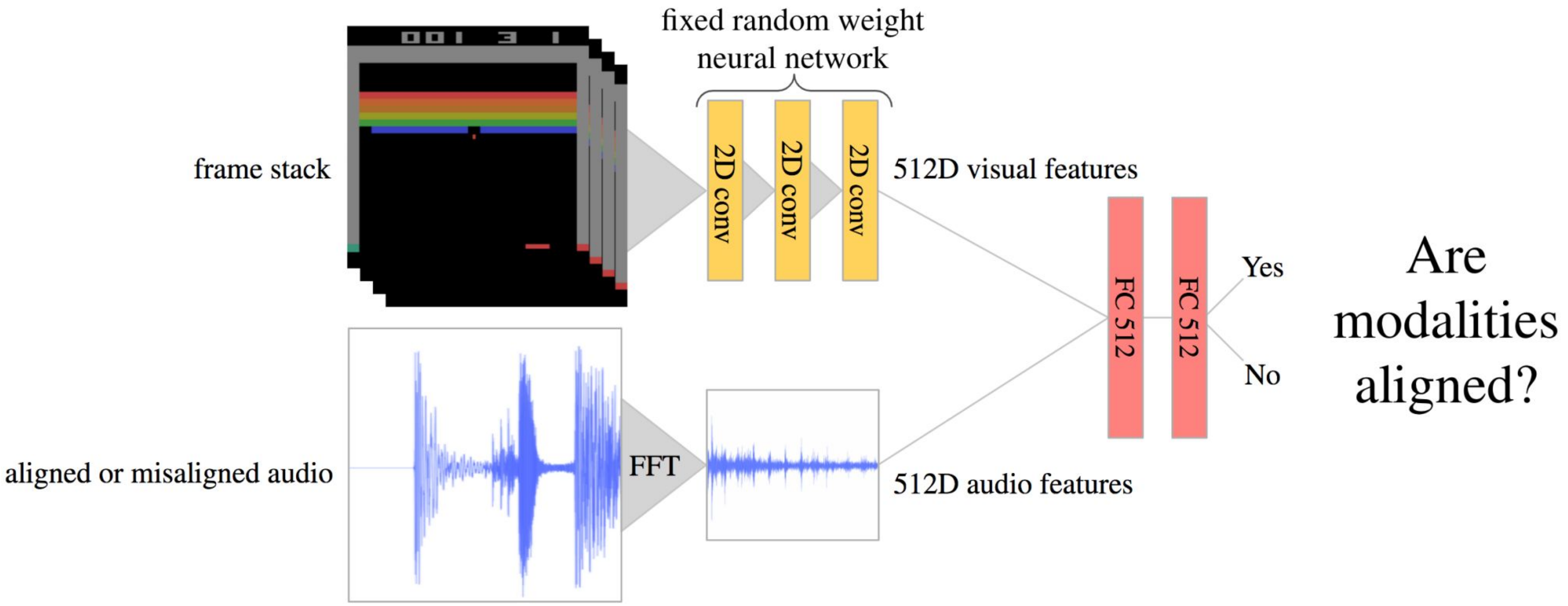
Victoria Dean, Shubham Tulsiani, Abhinav Gupta
NeurIPS 2020

Inspiration from humans

- Humans, especially babies, use multiple modalities to learn about the world
- Dember and Earl argue that intrinsic motivation comes from discrepancies between expected perception and actual stimulus

How can we exploit this for reinforcement learning?

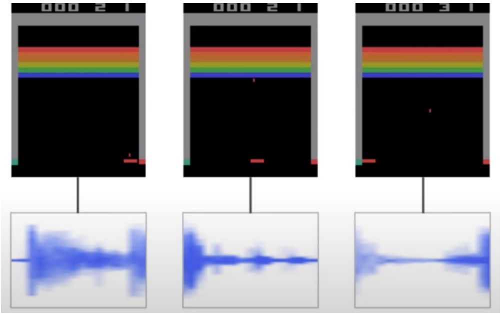
- Audio-video association discriminator



Victoria Dean, Shubham Tulsiani, Abhinav Gupta, "See, Hear, Explore: Curiosity via Audio-Visual Association" NeurIPS 2020

Data collection

- Agent policy is rolled out in parallel instances
- Trajectories from each instance are chunked into 128 time steps
- Time step consists of visual and sound features: $(v_t, s_t), t \in [1,128]$
  - Positive samples are matching pairs
  - Negative samples have true visual feature v_t and false sound feature s'_t
    - s'_t is uniformly sampled from the current trajectory



Victoria Dean, Shubham Tulsiani, Abhinav Gupta, "See, Hear, Explore: Curiosity via Audio-Visual Association" NeurIPS 2020

How do we train the discriminator?

- Weighted cross entropy loss

$$\mathcal{L}_t(v_t, s_t, z_t) = \begin{cases} -\log(D(v_t, s_t)), & \text{if } z_t = 1 \\ -\dfrac{||s_t - s'_t||_2}{\mathbb{E}_{\text{batch}}||s_t - s'_t||_2} \log(1 - D(v_t, s'_t)), & \text{if } z_t = 0 \end{cases}$$

- z_t is an indicator variable that is 1 when the true sound is used
- Weighting prevents punishment for similar false and true audio samples

Training the agent via intrinsic reward

- Intrinsic reward:

$$r_t^i := -\log(D(v_t, s_t))$$

- Policy is trained to maximize expected reward:

$$\max_\theta \mathbb{E}_{\pi(v_t;\theta)} \left[\sum_t \gamma^t r_t^i\right]$$

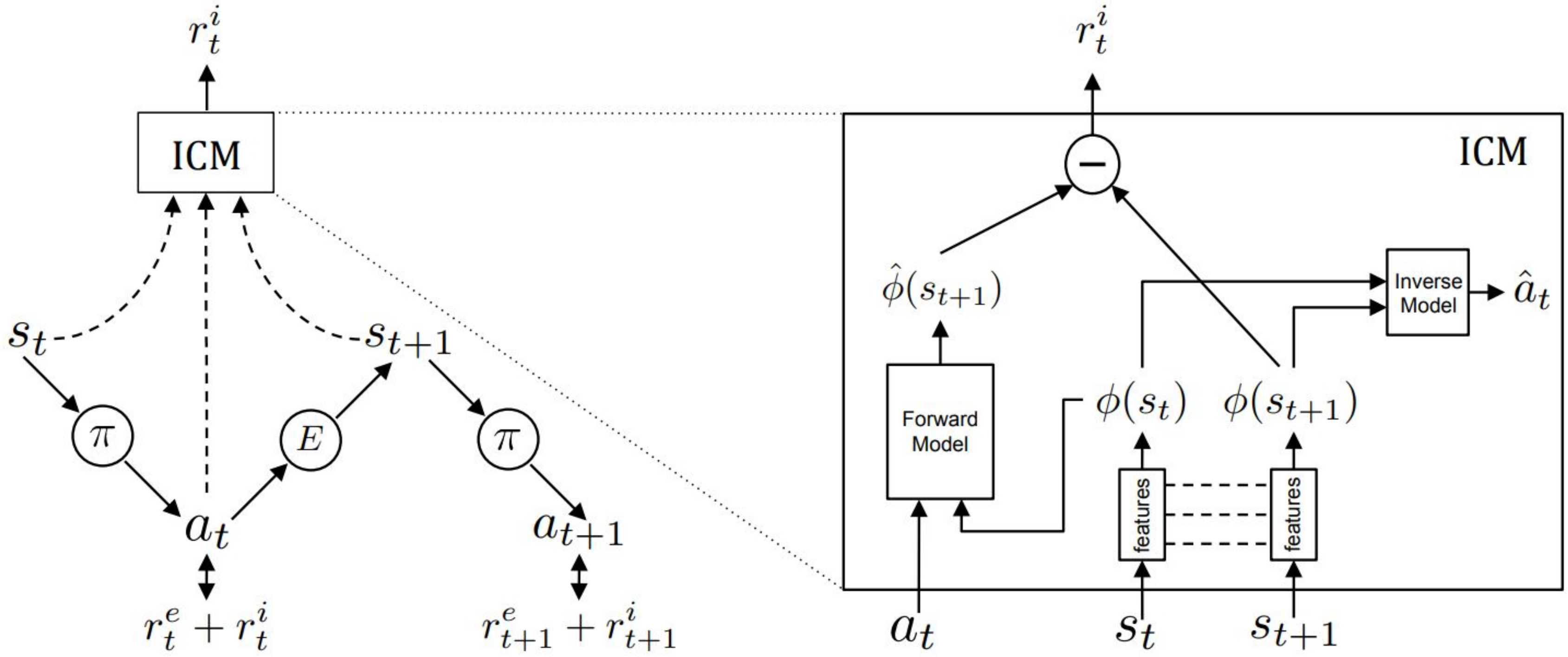
- Trained with a policy optimization technique, in this case PPO

Baselines

- Future prediction curiosity
- Exploration via disagreement
- Random network distillation (RND)

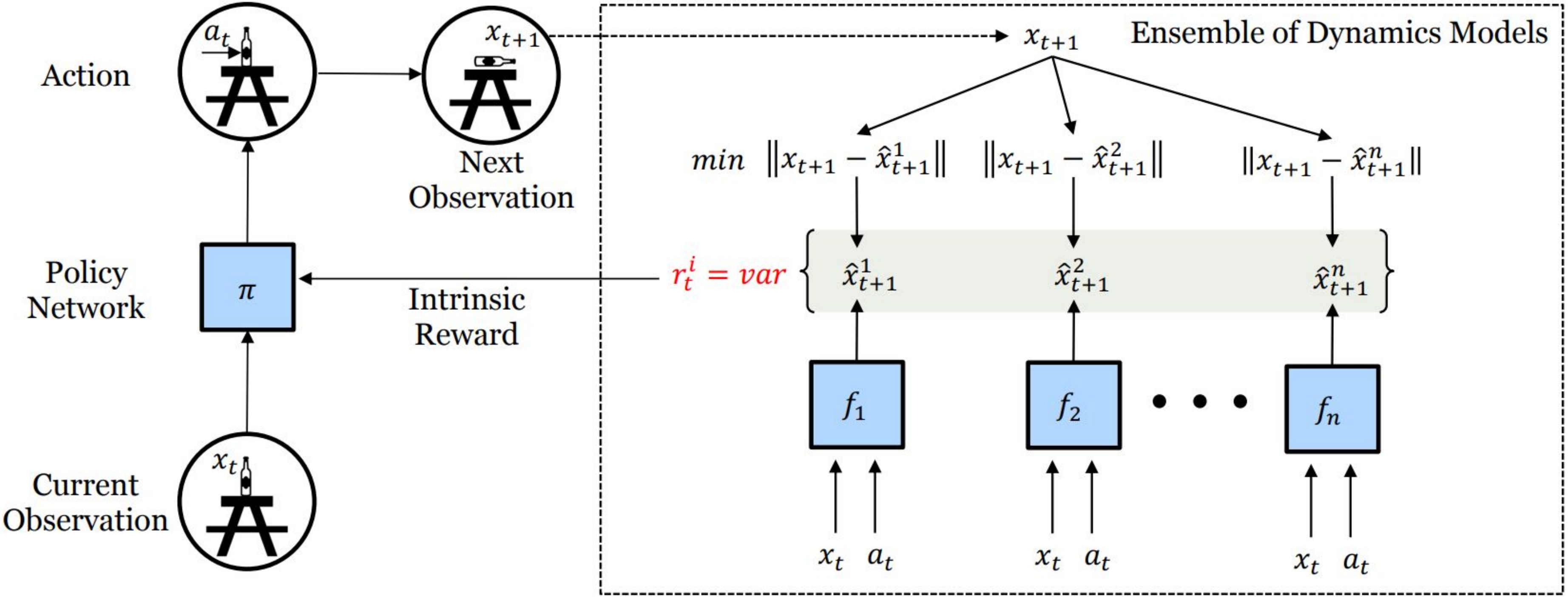
- Hyperparameters for policy learning are the same across all approaches
- CNN features are random for all approaches

Future prediction curiosity



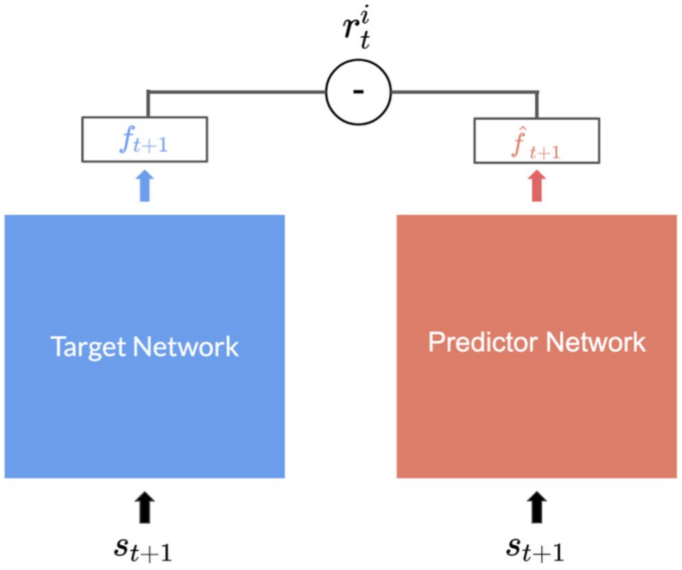
Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, Trevor Darrell, "Curiosity-driven Exploration by Self-supervised Prediction" ICML 2020

Exploration via disagreement



Deepak Pathak, Dhiraj Gandhi, Abhinav Gupta, "Self-Supervised Exploration via Disagreement" ICML 2019

Random network distillation



Target network is randomly initialized

Target network will output a fixed feature representation of s_{t+1}

Predictor Network will tries to predict the target network's output $\hat{f}_{t+1}$

$$r_t^i = \left\| \underbrace{\hat{f}\left(s_{t+1}\right)}_{\text{Prediction of the Feature Representation of next state}} - \underbrace{f\left(s_{t+1}\right)}_{\text{Target Feature Representation of next state}} \right\|_{2}^{2}$$

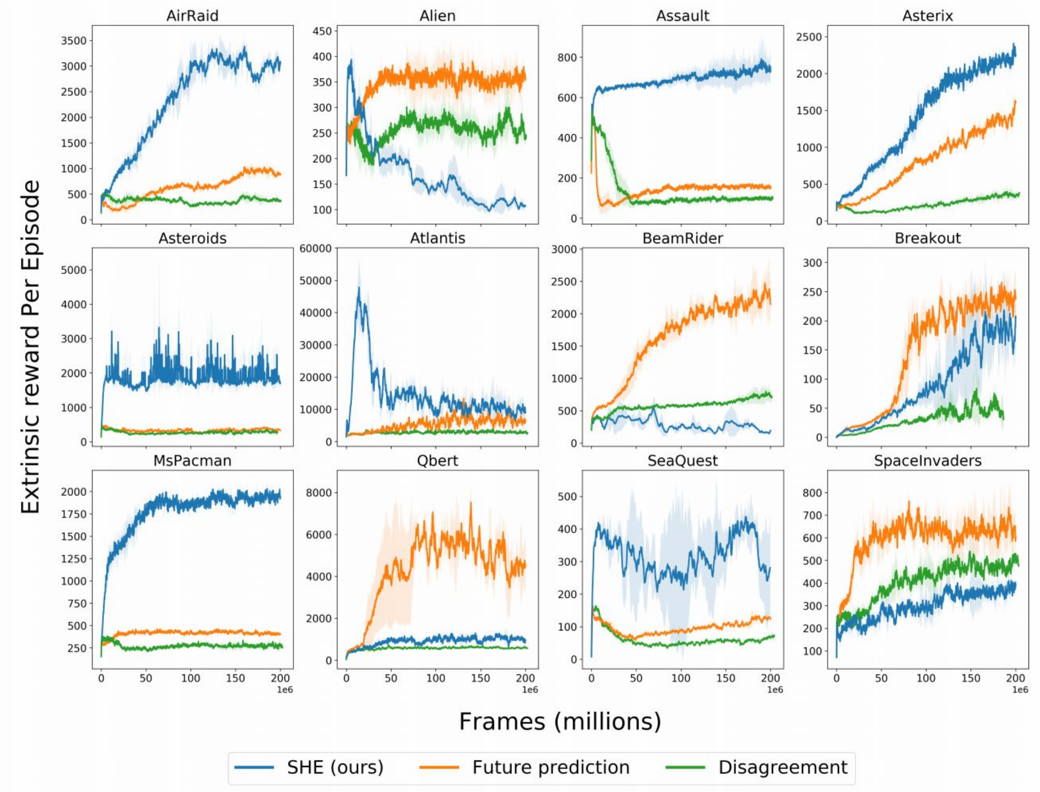
L2 norm (to output a scalar)

https://medium.com/data-from-the-trenches/curiosity-driven-learning-through-random-network-distillation-488ffd8e5938

Evaluation environment - Atari

- Evaluated on 12 Atari games
  - Some games excluded due to no audio (e.g. Amidar, Pong)
  - Other games excluded due to background music (e.g. RoadRunner, Super Mario Bros)
- Trained for 200 million frames (allegedly more sample efficient)

Results - Atari training curves



Victoria Dean, Shubham Tulsiani, Abhinav Gupta, "See, Hear, Explore: Curiosity via Audio-Visual Association" NeurIPS 2020

Failure case - trivial audio-visual association

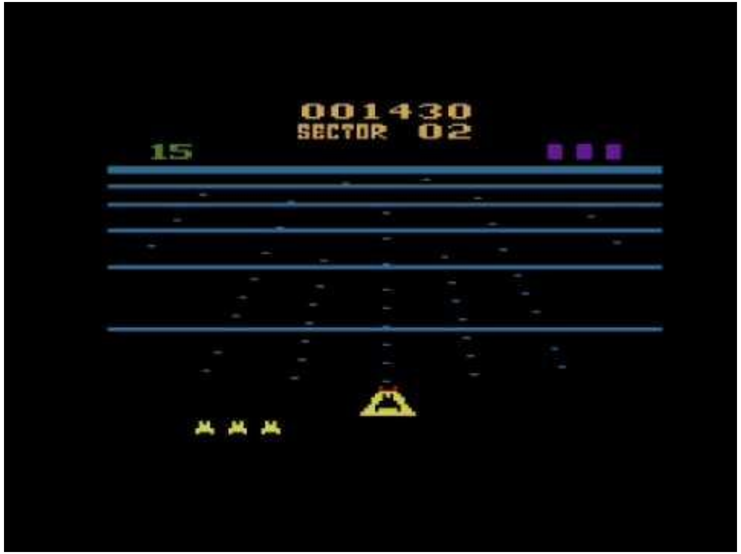
- Easy discriminator task leads to low agent rewards
- Visiting already-learned states necessary for high extrinsic reward

Qbert

Atlantis

Failure case - repetitive background sounds

- Difficult to visually associate sounds
- Trouble learning basic cases makes agent unmotivated to explore



BeamRider

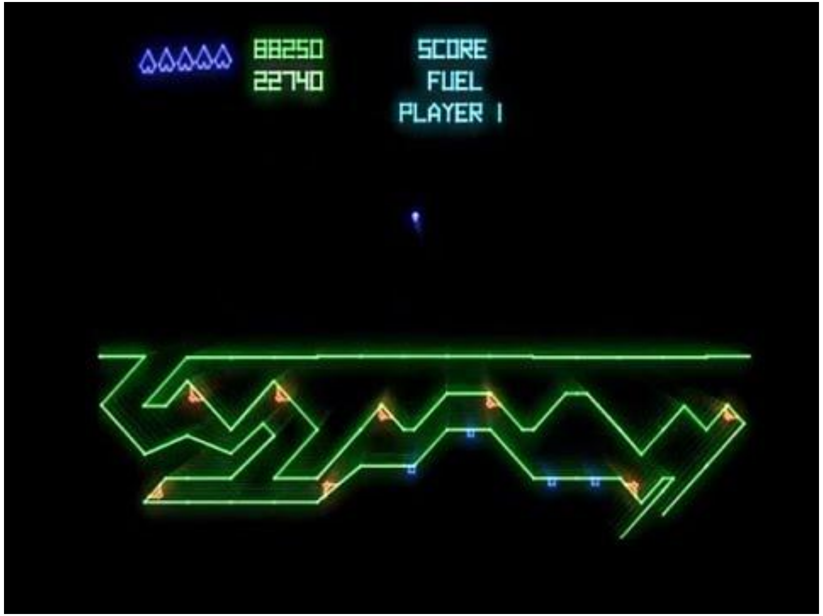
Space Invaders

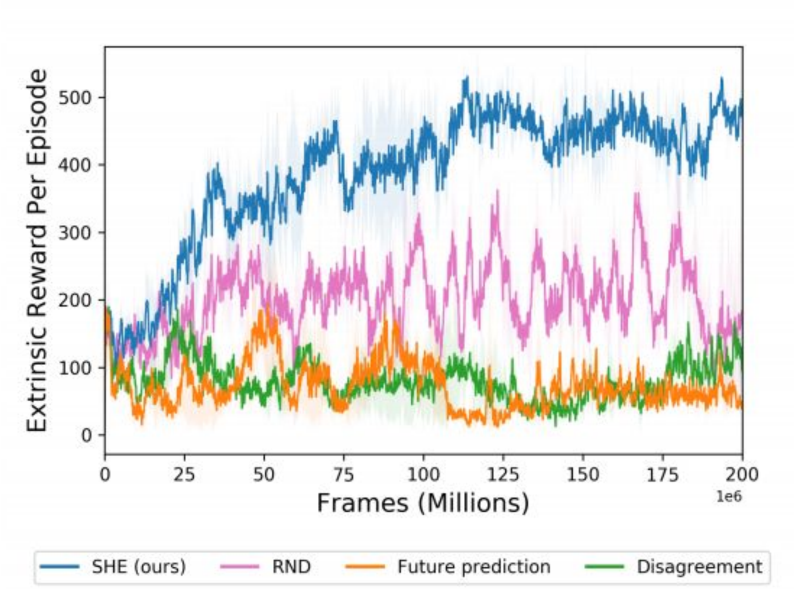
Failure case - learned repetitive sounds?

- Agent gets stuck in a loop of passing from one side of the screen to the other in Alien
- Slight delay in sound makes alignment difficult

Success case - Gravitar

- Hard exploration environment
- Visual dynamics not very interesting - audio-visual associations are



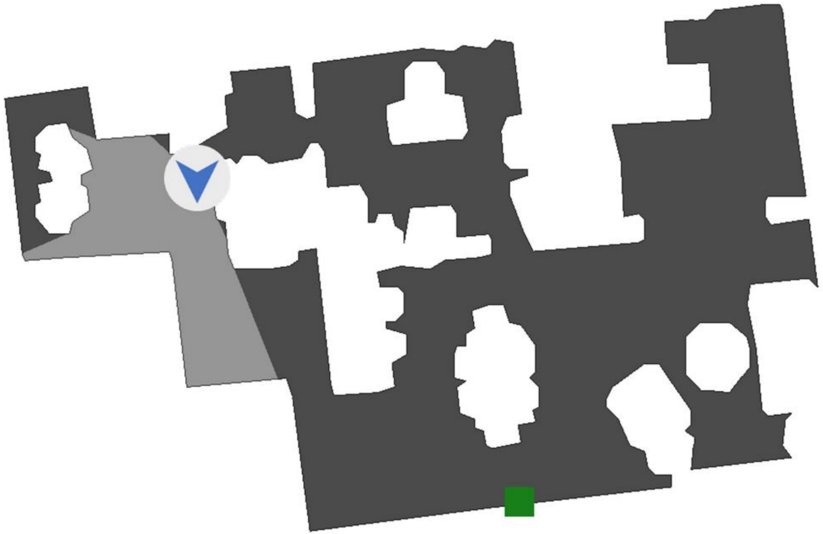


Victoria Dean, Shubham Tulsiani, Abhinav Gupta, "See, Hear, Explore: Curiosity via Audio-Visual Association" NeurIPS 2020

Evaluation environment - Habitat

- Photorealistic simulator using Replica Dataset
- Sound source emits a fixed audio clip less than one second long

Apartment 0 render

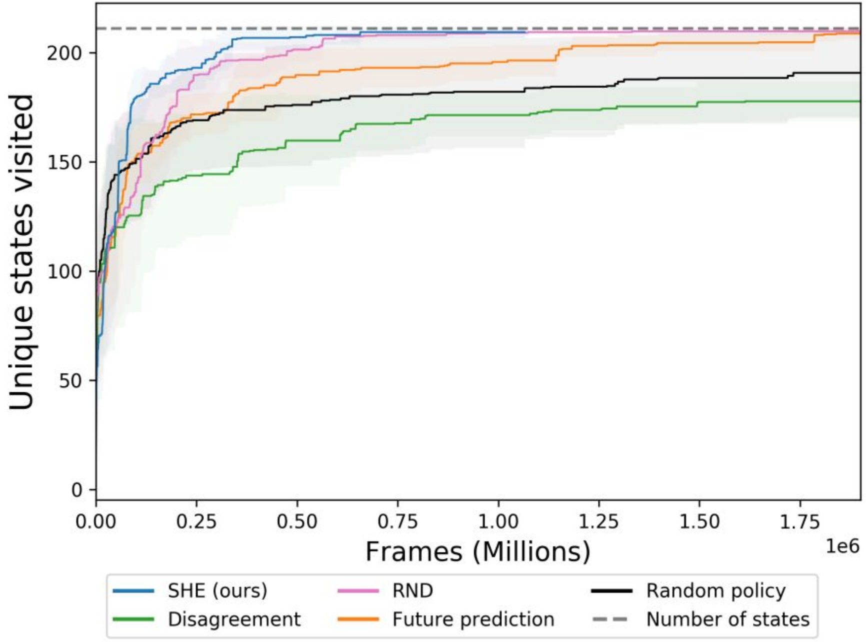


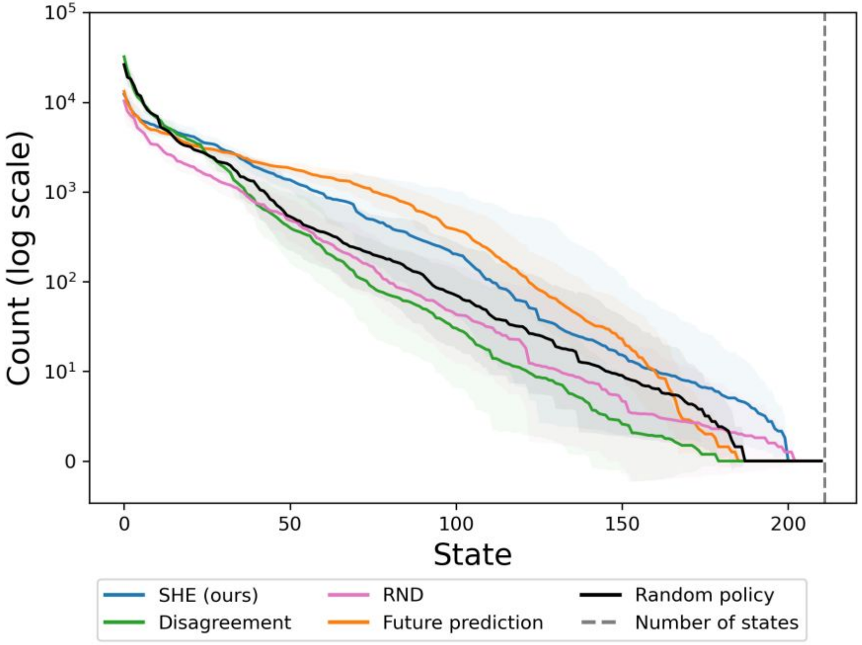
Apartment 0 bird's eye view

https://github.com/facebookresearch/Replica-Dataset
Victoria Dean, Shubham Tulsiani, Abhinav Gupta, "See, Hear, Explore: Curiosity via Audio-Visual Association" NeurIPS 2020

Habitat - results

- Authors claim significant gains over baselines

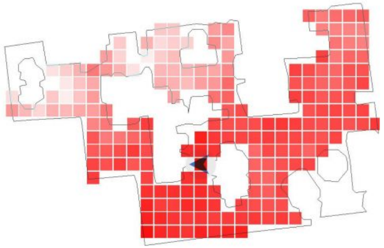




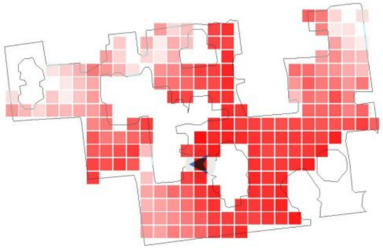
Victoria Dean, Shubham Tulsiani, Abhinav Gupta, "See, Hear, Explore: Curiosity via Audio-Visual Association" NeurIPS 2020

Habitat - results cont’d

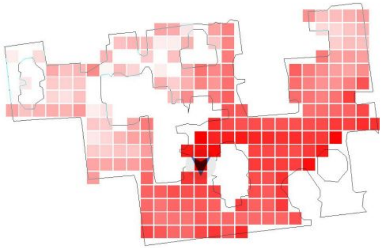
- Heatmaps do not seem to show particularly superior performance
- Agents start facing different directions?



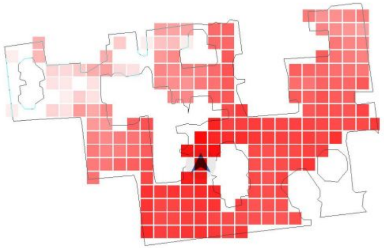
(a) SHE (ours)



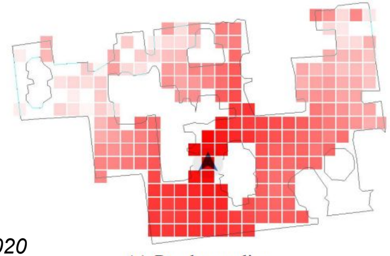
(b) Future prediction



(c) Disagreement



(d) RND

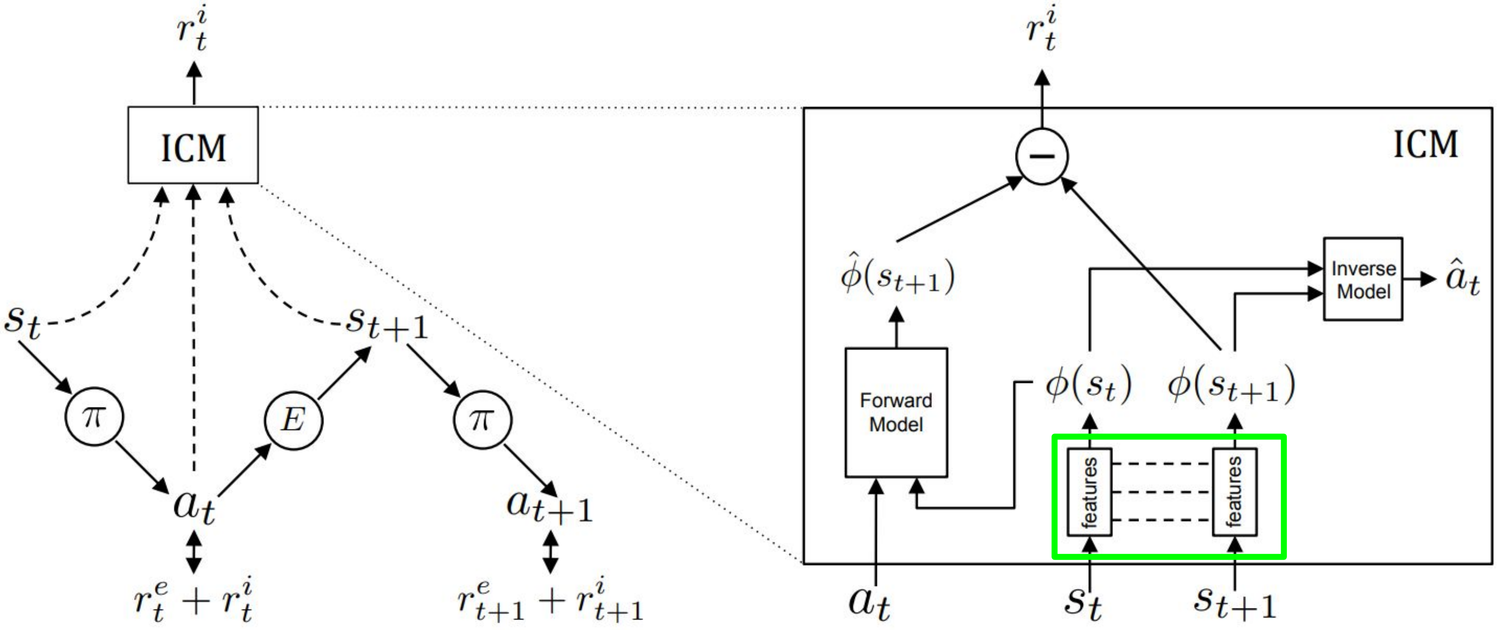


(e) Random policy

Victoria Dean, Shubham Tulsiani, Abhinav Gupta, “See, Hear, Explore: Curiosity via Audio-Visual Association” NeurIPS 2020

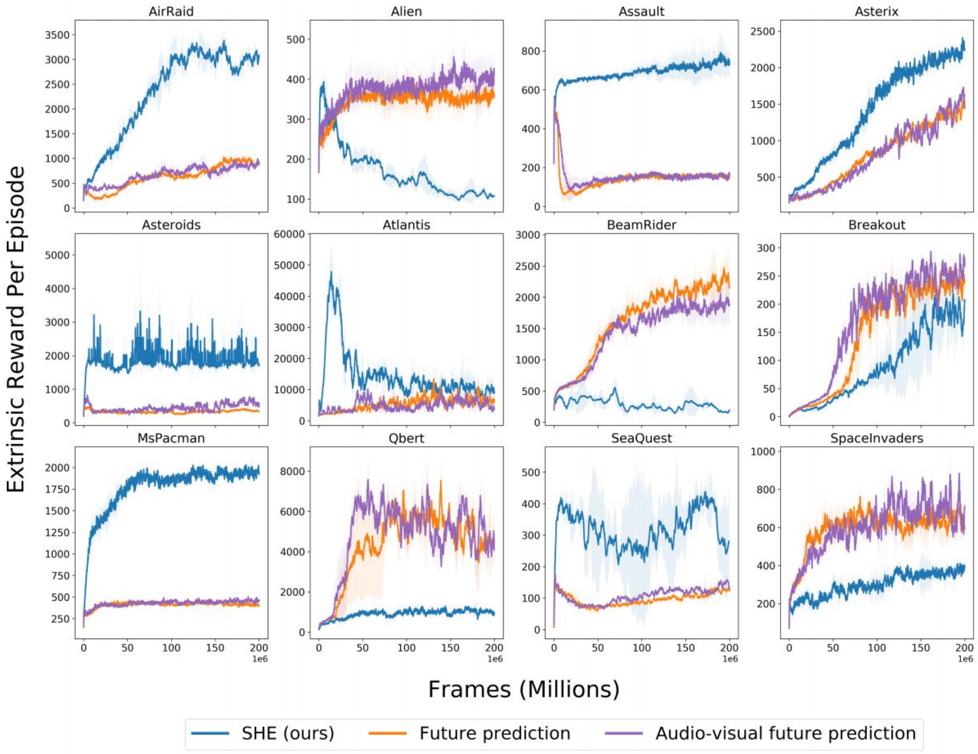
Ablations - future prediction with audio

- Concatenate audio features to visual features



Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, Trevor Darrell, “Curiosity-driven Exploration by Self-supervised Prediction” ICML 2020

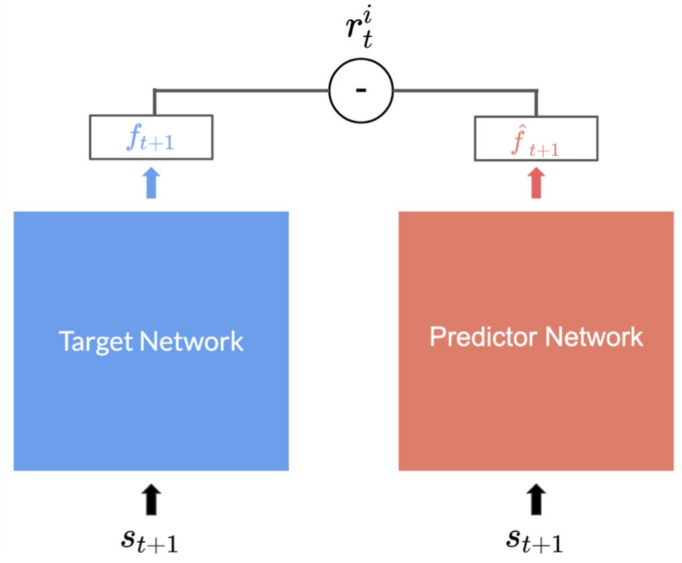
Results - future prediction with audio



Victoria Dean, Shubham Tulsiani, Abhinav Gupta, "See, Hear, Explore: Curiosity via Audio-Visual Association" NeurIPS 2020

Ablations - RND with audio

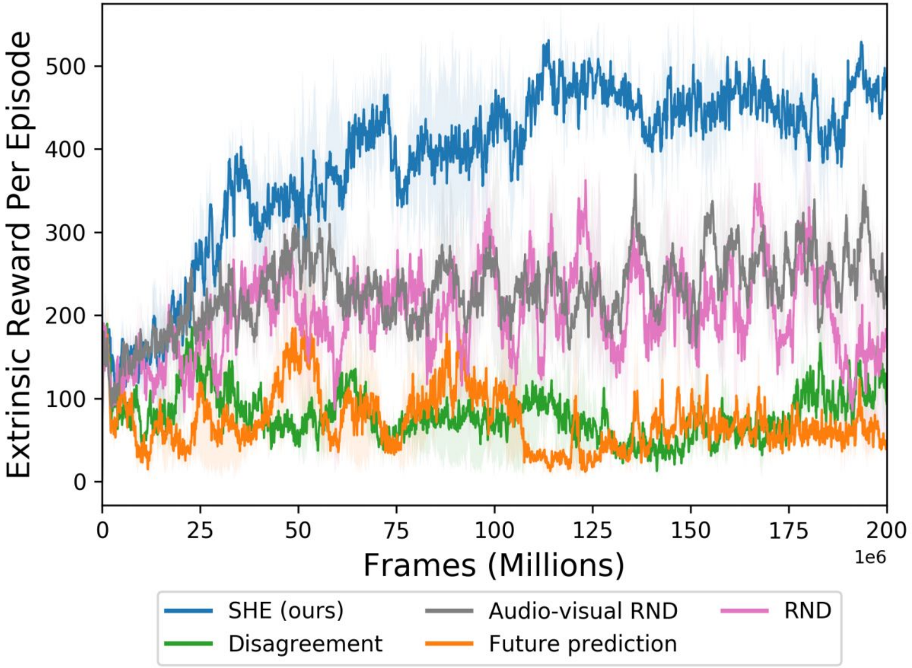
- Image and audio are converted to features with convolutional and dense layers respectively, then concatenated



https://medium.com/data-from-the-trenches/curiosity-driven-learning-through-random-network-distillation-488ffd8e5938

Results - RND with audio

- Authors note that differing sparsities between video and audio features makes this difficult
- Claim their method is better because it doesn't need tuning

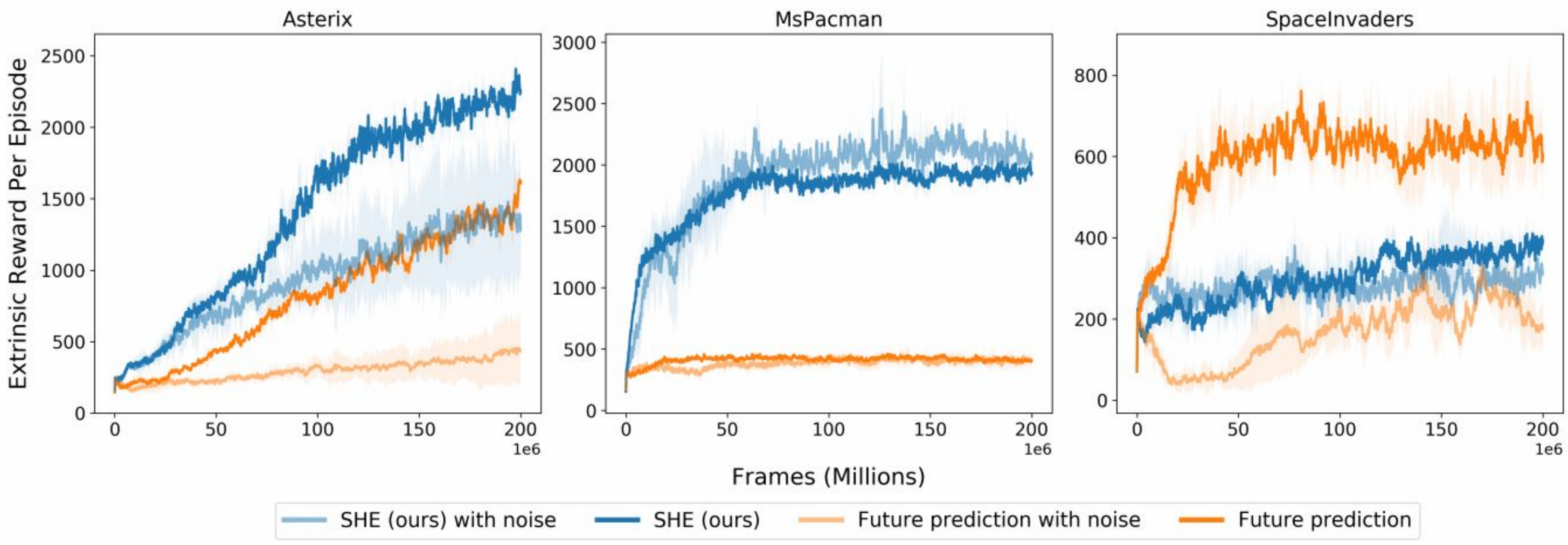


Audio-visual RND on Gravitar

Victoria Dean, Shubham Tulsiani, Abhinav Gupta, "See, Hear, Explore: Curiosity via Audio-Visual Association" NeurIPS 2020

Ablations - robustness to noise

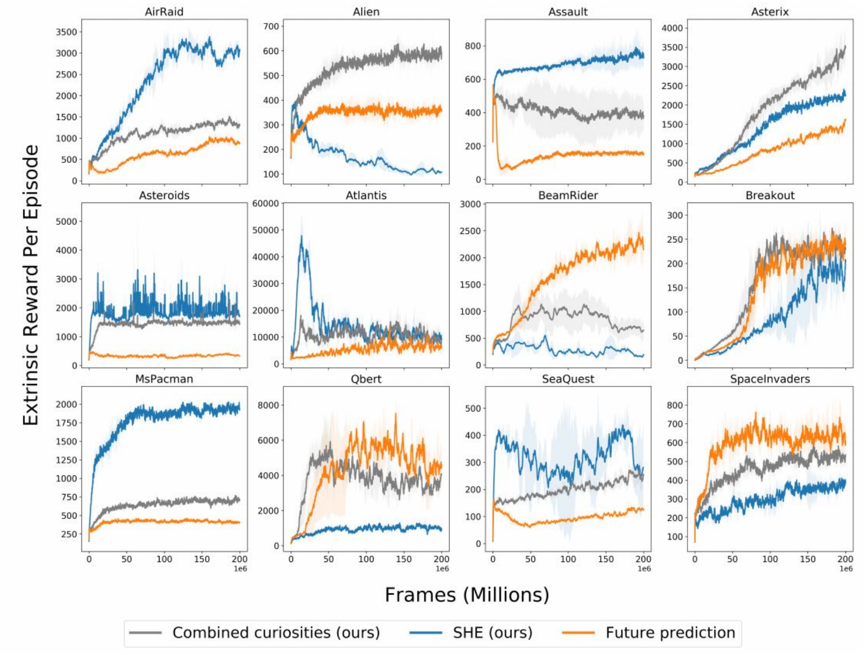
- Gaussian noise added to audio and visual feature vector inputs



Victoria Dean, Shubham Tulsiani, Abhinav Gupta, "See, Hear, Explore: Curiosity via Audio-Visual Association" NeurIPS 2020

Ablations - multiple curiosity modules

- Sum the rewards from future prediction and audio-visual discriminator



Victoria Dean, Shubham Tulsiani, Abhinav Gupta, "See, Hear, Explore: Curiosity via Audio-Visual Association" NeurIPS 2020

Final thoughts

Pros:

- Work is interesting - a successful implementation of multi-modal curiosity
- Shows strong performance on certain Atari games
  - Performs well on some challenging games like Gravitar

Cons:

- Habitat experiment does not seem particularly convincing
- Audio ablation does not seem totally fair
- Method has lots of limitations - no sound, too much sound, etc.
- Performs significantly worse on some Atari games with more information

Discussion

- What other modalities might provide useful information for exploration?

- Ideally, adding additional information does not degrade performance below previous systems. How can we incorporate sound into a reinforcement learning system without degrading performance?

- How else might curiosity be instilled into reinforcement learning systems?