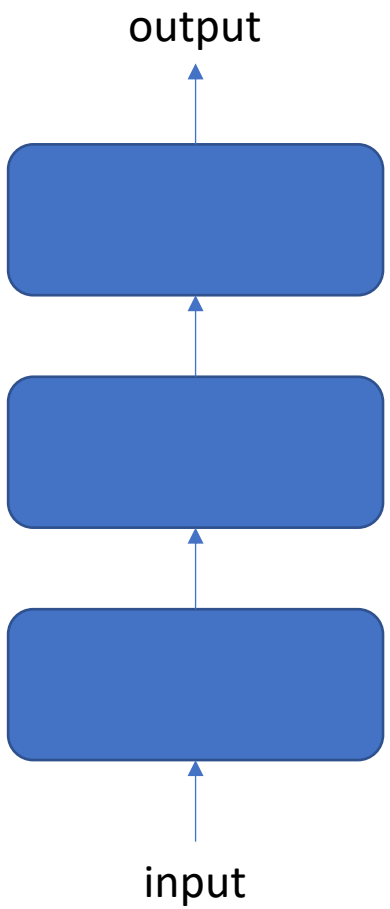
Progressive Neural Networks

Чистяков Глеб, гр. 162

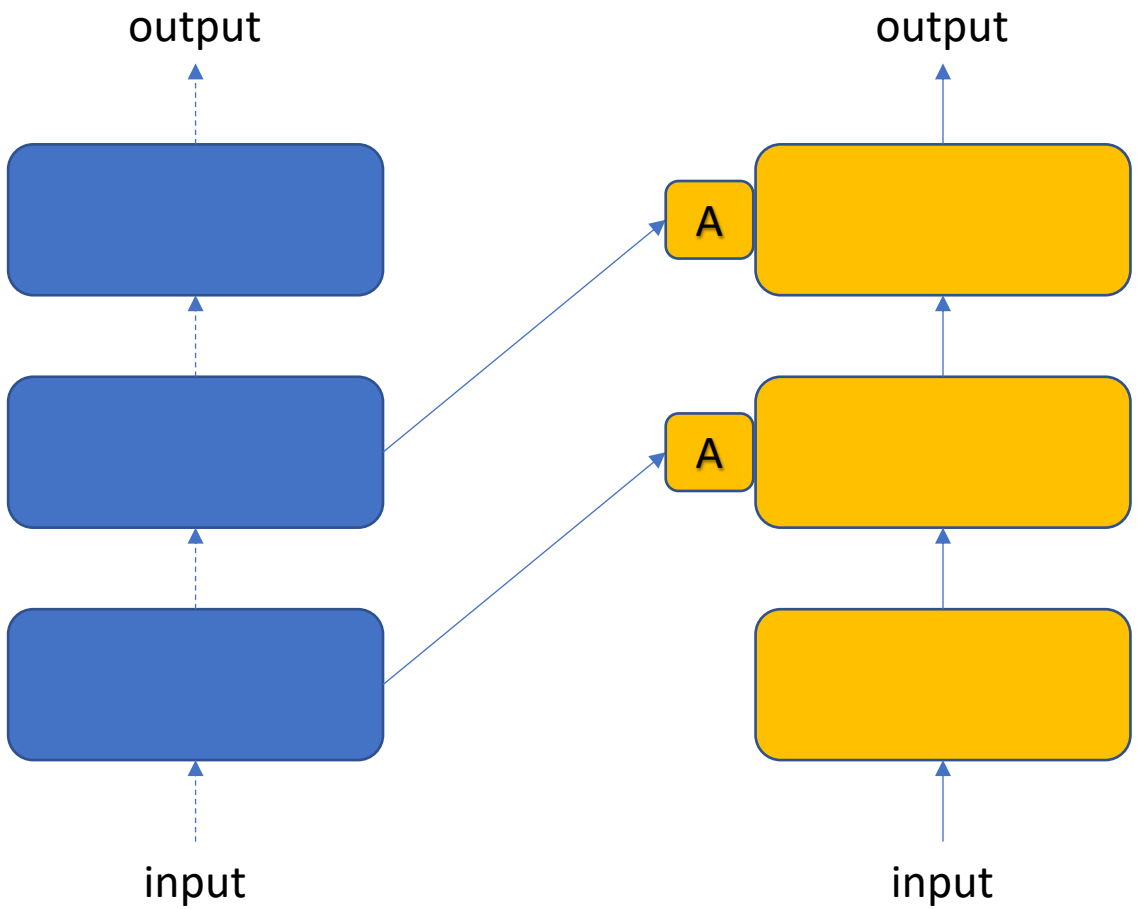
Progressive Networks

- Способность включать предварительные знания на каждом уровне иерархии объектов
- Способность повторно использовать старые вычисления и изучать новые
- Невосприимчивость к катастрофическому забвению (catastrophic forgetting)

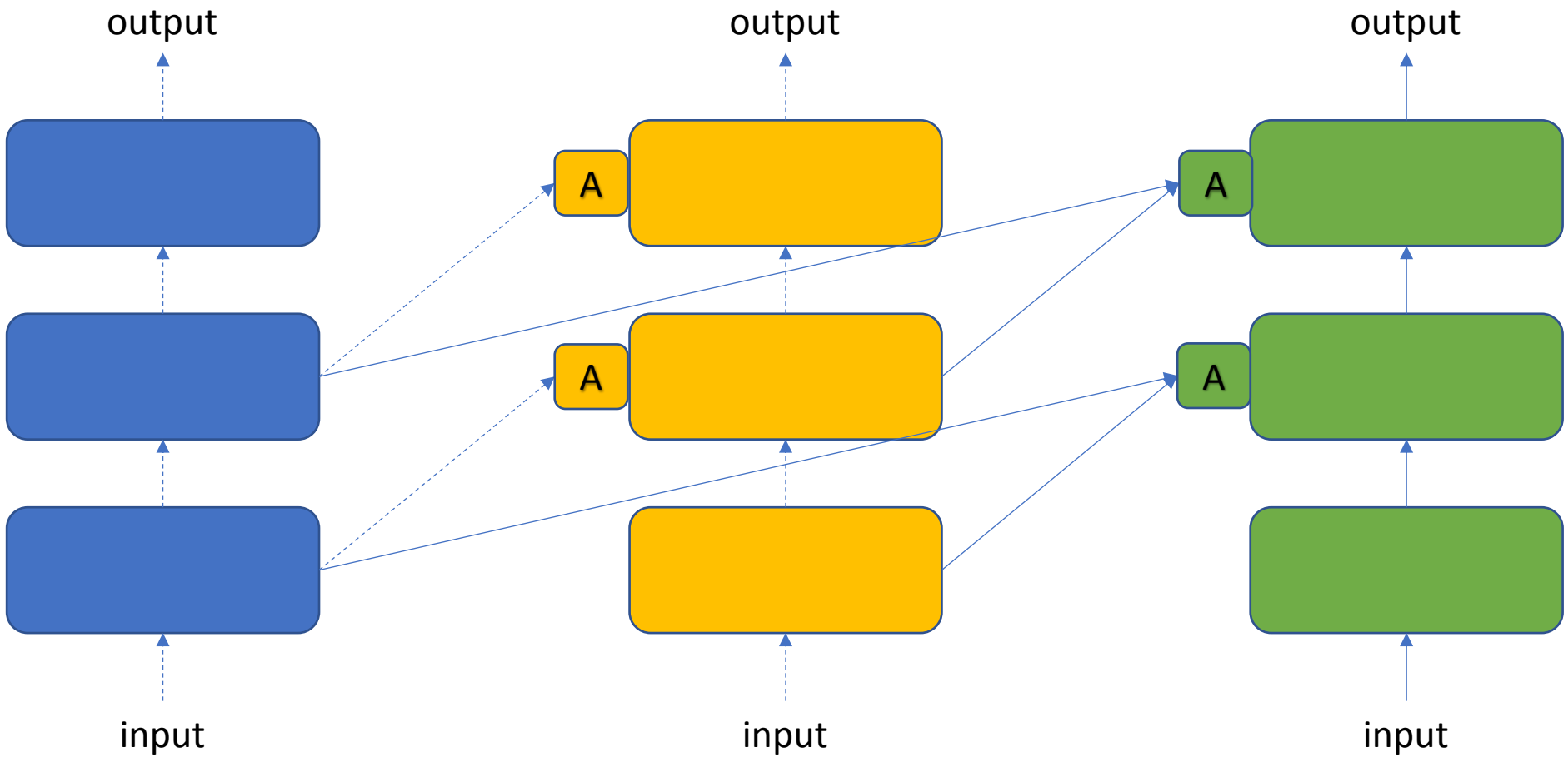
Progressive Networks



Progressive Networks



Progressive Networks



Hidden activations

$$h_i^{(k)} = f\left(W_i^{(k)} h_{i-1}^{(k)} + \sum_{j<k} U_i^{(k:j)} h_{i-1}^{(j)}\right)$$

$W_i^{(k)} \in \mathbb{R}^{n_i \times n_{i-1}}$ – матрица весов слоя i столбца k

$U_i^{(k:j)} \in \mathbb{R}^{n_i \times n_j}$ – боковые соединения от слоя $i - 1$ столбца j до слоя $i - 1$ столбца k

$\mathrm{f(x)} = \max(0, \mathrm{x})$

Adapters

$h_{i-1}^{(<k)} = [h_{i-1}^{(1)} \dots h_{i-1}^{(j)} \dots h_{i-1}^{(k-1)}]$ – вектор предыдущих параметров

$$h_i^{(k)} = \sigma\left(W_i^{(k)} h_{i-1}^{(k)} + U_i^{(k:j)} \sigma(V_i^{(k:j)} \alpha_{i-1}^{(<k)} h_{i-1}^{(<k)})\right)$$

$V_i^{(k:j)} \in \mathbb{R}^{n_{i-1} \times n_{i-1}^{(<k)}}$ – матрица проекции

Transfer Analysis

- Average Perturbation Sensitivity (APS)
- Average Fisher Sensitivity (AFS)

Transfer Analysis

Average Perturbation Sensitivity (APS)

$\Lambda_i^{(k)} = \frac{1}{\sigma_i^{2(k)}}$ – точность шума, вводимого в слой i столбца k

$$\mathrm{APS}(i,k) = \frac{\Lambda_i^{(k)}}{\sum_k \Lambda_i^{(k)}}$$ - оценка чувствительности слоя *i* столбца *k*

Transfer Analysis

Average Fisher Sensitivity (AFS)

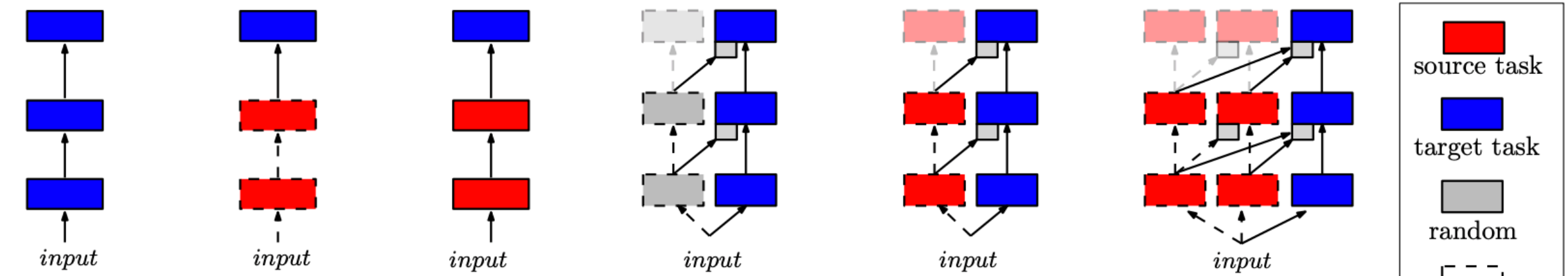
$$\hat{F}_i^{(k)} = \mathbb{E}_{\rho(s,a)} \left[\frac{\partial \log \pi}{\partial \hat{h}_i^{(k)}} \frac{\partial \log \pi}{\partial \hat{h}_i^{(k)}}^T \right]$$ - матрица Фишера

$$\pi^{(k)}(a \mid s) := h_L^{(k)}(s)$$ - Политика *k*-го столбца, принимающая состояние окружения и выдающая вероятность над действиями

$$\mathbf{AFS}(i,k,m) = \frac{\hat{F}_i^{(k)}(m,m)}{\sum_k \hat{F}_i^{(k)}(m,m)}$$ - средняя чувствительность Фишера признака *m* слоя *i* столбца *k*

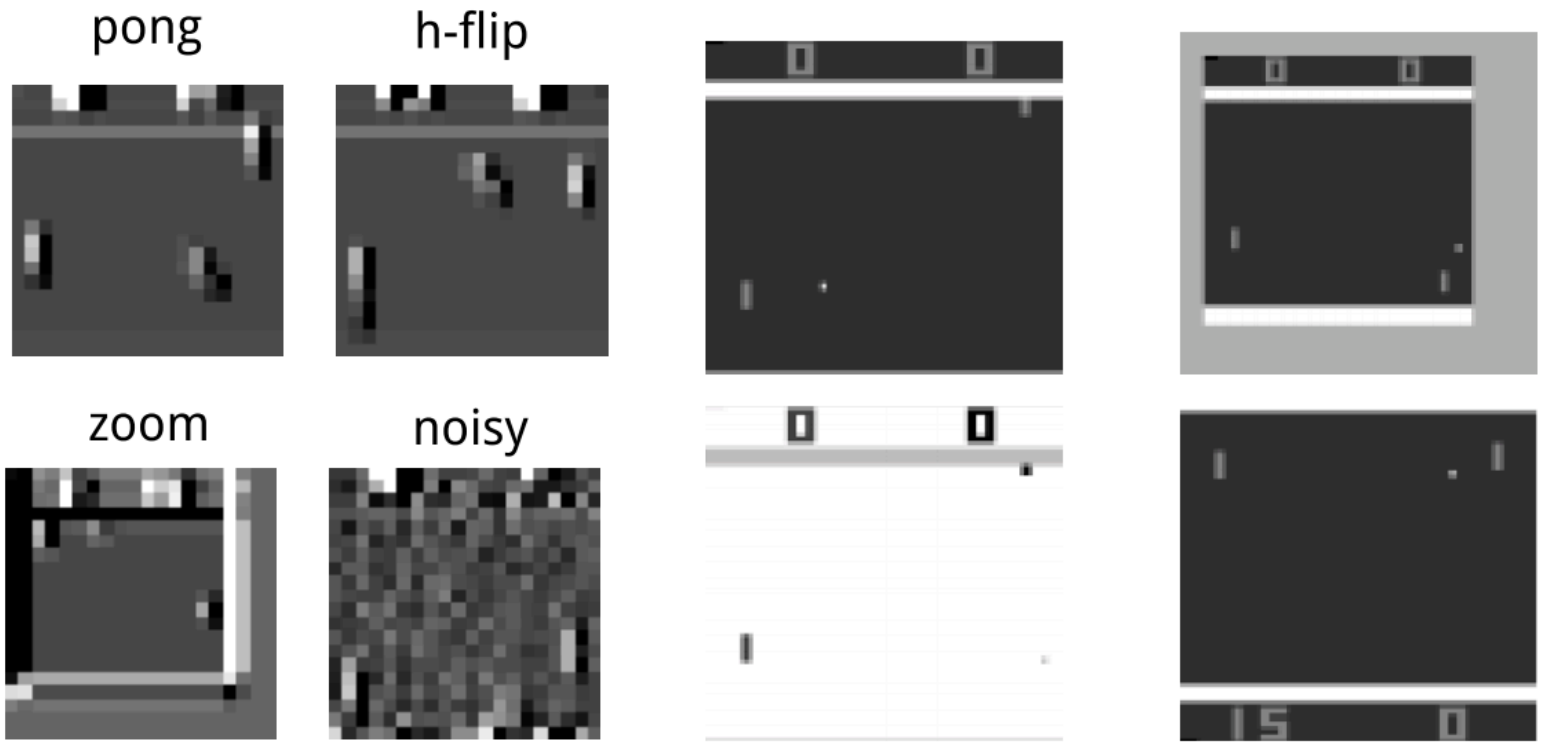
$$\mathbf{AFS}(i,k) = \sum_m \mathbf{AFS}(i,k,m)$$ - оценка чувствительности слоя *i* столбца *k*

Experiments



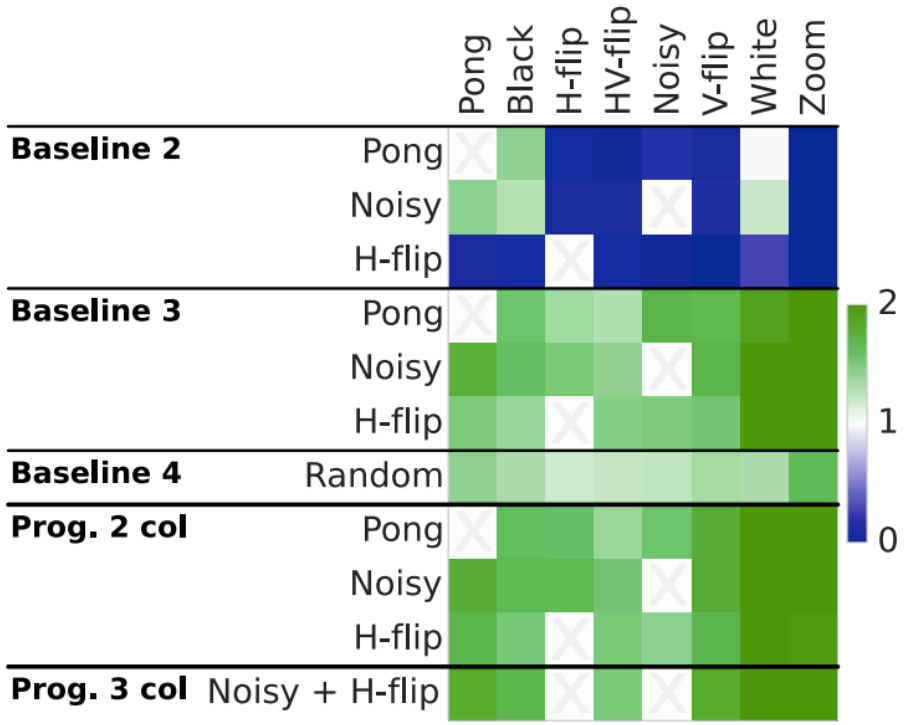
(1) **Baseline 1** (2) **Baseline 2** (3) **Baseline 3** (4) **Baseline 4** (5) **Progressive Net 2 columns** (6) **Progressive Net 3 columns**

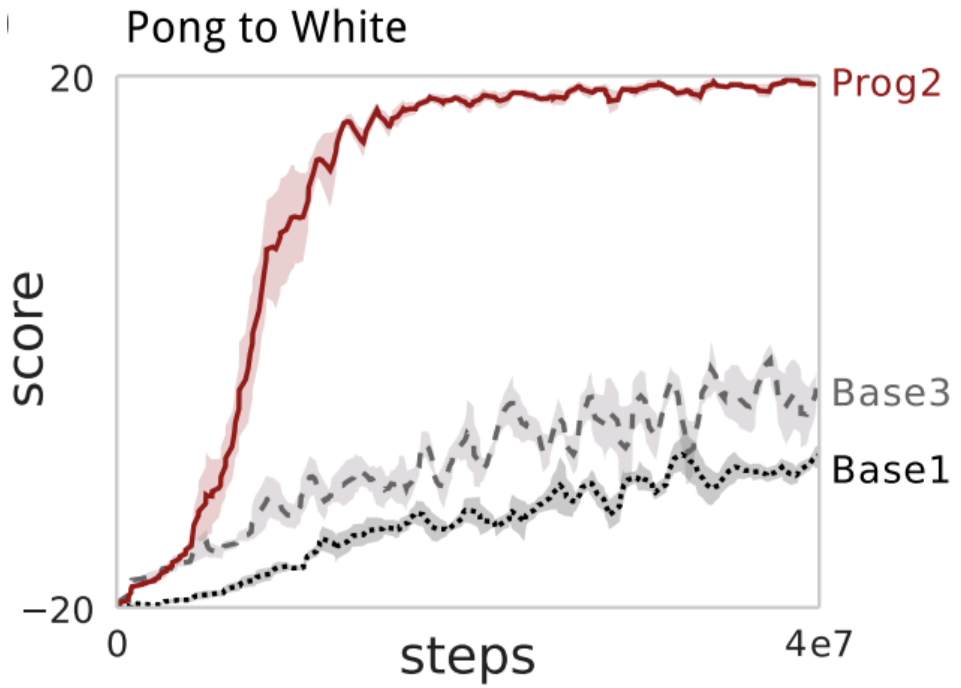
Pong Soup

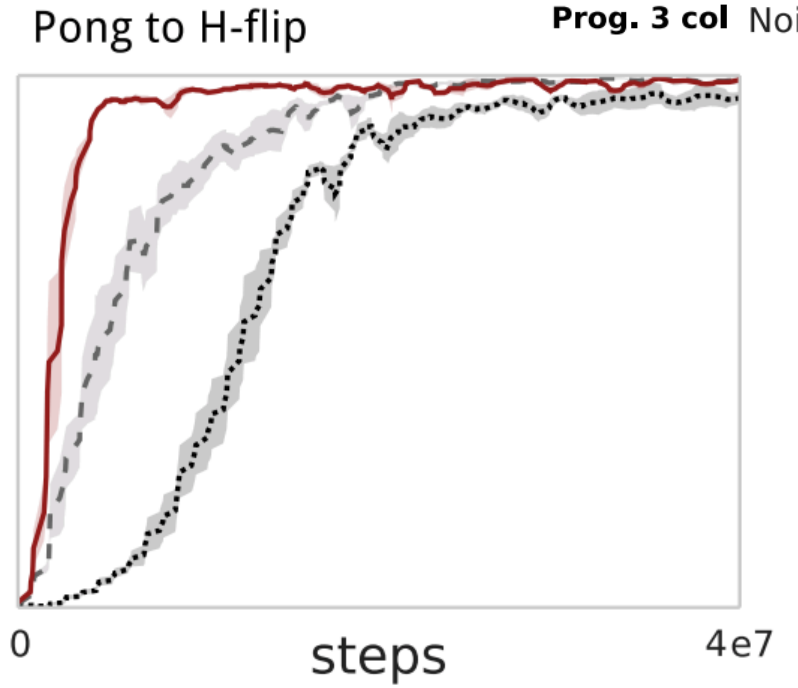


- Noisy (frozen Gaussian noise is added to the inputs)
- Black (black background)
- White (white background)
- Zoom (input is scaled by 75% and translated)
- V-flip (input is vertically flipped)
- H-flip (input is horizontally flipped)
- VH-flip (input is horizontally and vertically flipped)

Pong Soup


Pong
Black
H-flip
HV-flip
Noisy
V-flip
White
Zoom
Baseline 2
Pong
Noisy
H-flip
Baseline 3
Pong
Noisy
H-flip
Baseline 4
Random
Prog. 2 col
Pong
Noisy
H-flip
Prog. 3 col
Noisy + H-flip
2
1
0


Pong to White
20
−20
score
0
steps
4e7
Prog2
Base3
Base1


Pong to H-flip
0
steps
4e7

Atari games

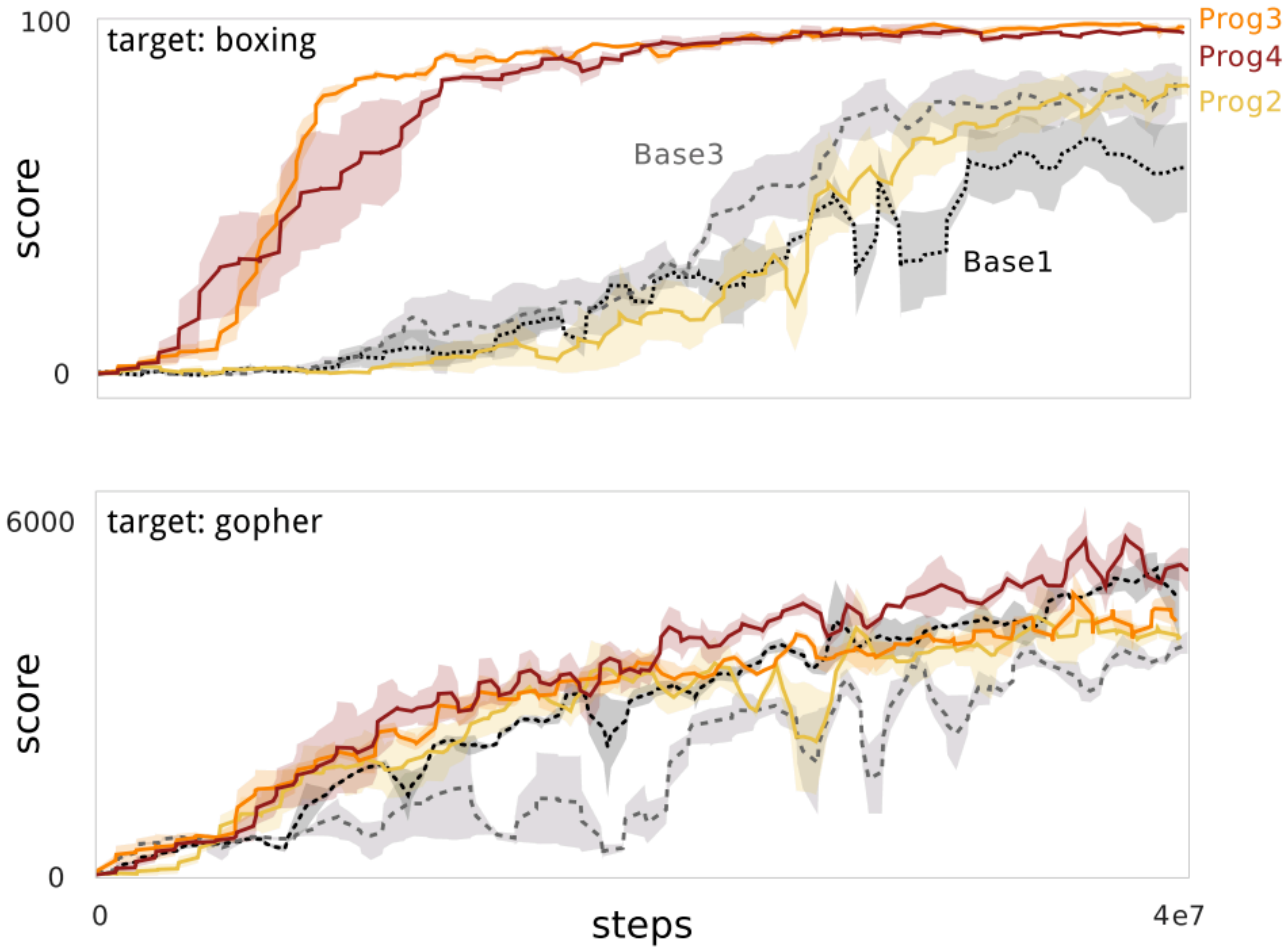
Source games:

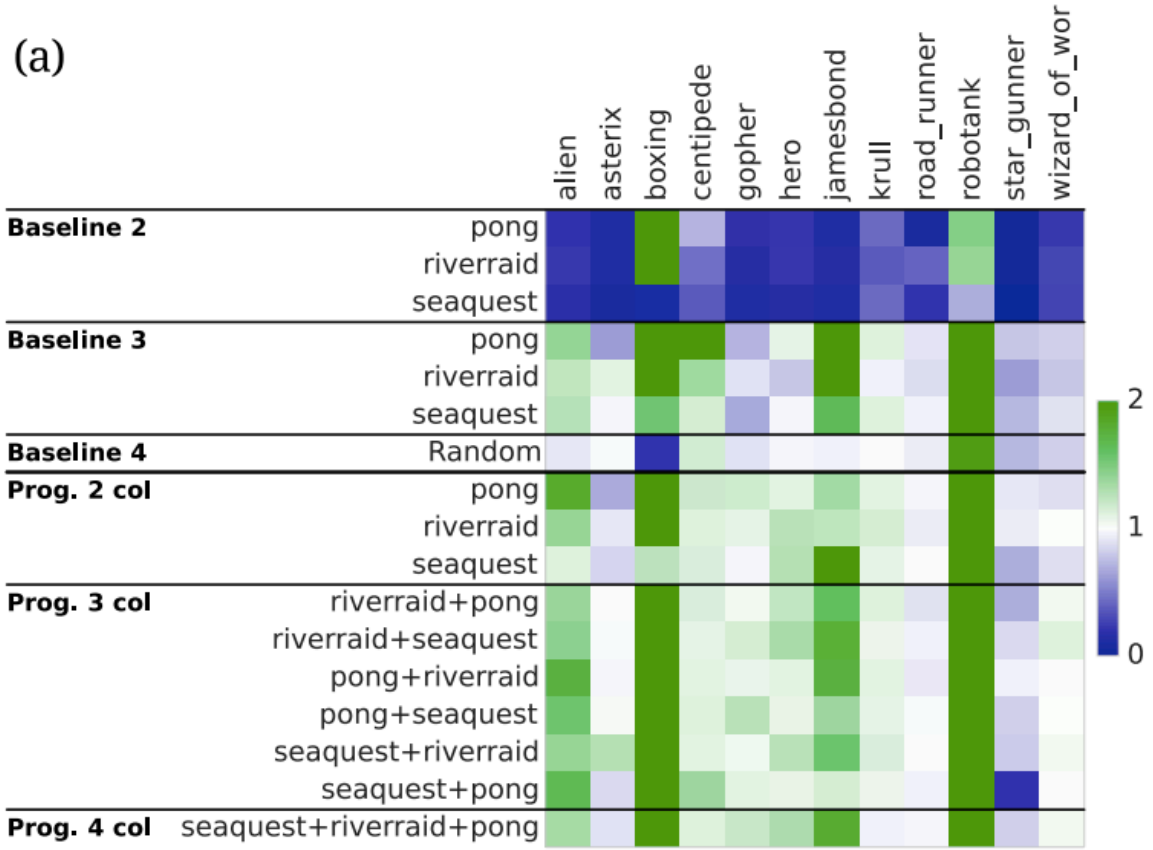
- Pong
- River Raid
- Seaquest

Target games:

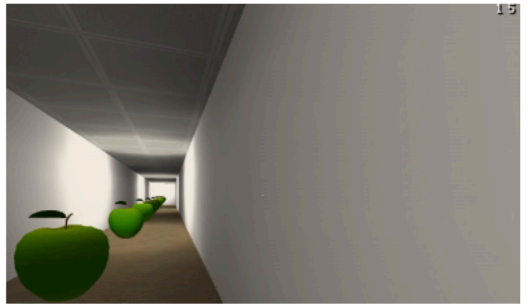
- Alien
- Asterix
- Boxing
- Centipede
- Gopher
- Hero
- James Bond
- Krull Robotank
- Road Runner
- Star Gunner
- Wizard of Wor

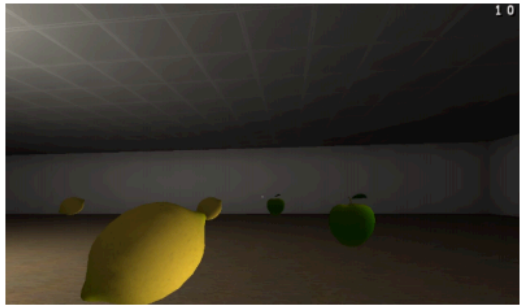
Atari games


target: boxing
100
0
score
Prog3
Prog4
Prog2
Base3
Base1
target: gopher
6000
0
score
0
steps
4e7


(a)
alien
asterix
boxing
centipede
gopher
hero
jamesbond
krull
road_runner
robotank
star_gunner
wizard_of_wor
Baseline 2
pong
riverraid
seaquest
Baseline 3
pong
riverraid
seaquest
Baseline 4
Random
Prog. 2 col
pong
riverraid
seaquest
Prog. 3 col
riverraid+pong
riverraid+seaquest
pong+riverraid
pong+seaquest
seaquest+riverraid
seaquest+pong
Prog. 4 col
seaquest+riverraid+pong
2
1
0

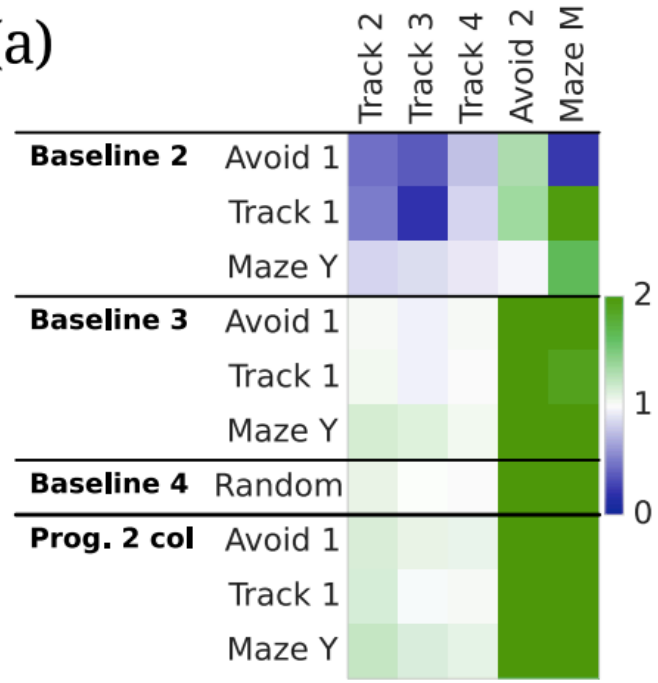
Labyrinth games

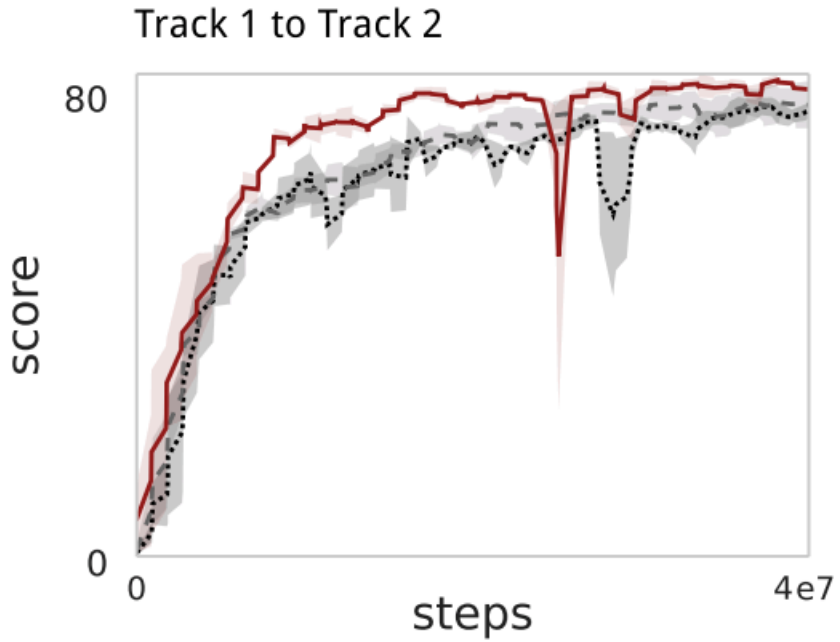


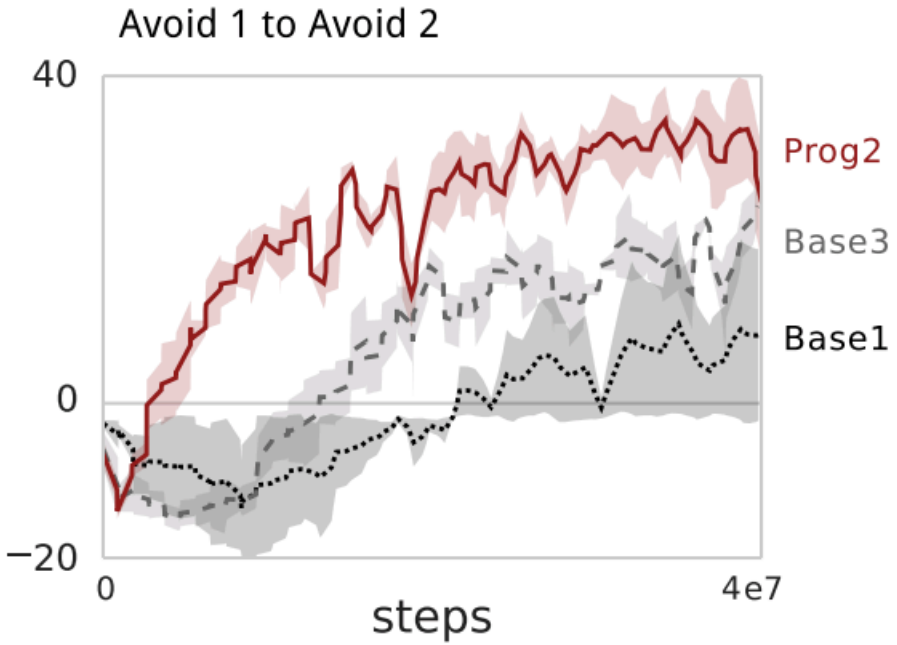


- Seek Track 1: simple corridor with many apples
- Seek Track 2: U-shaped corridor with many strawberries
- Seek Track 3: Ω-shaped, with 90o turns, with few apples
- Seek Track 4: Ω-shaped, with 45o turns, with few apples
- Seek Avoid 1: large square room with apples and lemons
- Seek Avoid 2: large square room with apples and mushrooms
- Seek Maze M : M-shaped maze, with apples at dead-ends
- Seek Maze Y : Y-shaped maze, with apples at dead-ends

Labyrinth games







Experiments

	Pong Soup		Atari		Labyrinth	
	Mean (%)	Median (%)	Mean (%)	Median (%)	Mean (%)	Median (%)
Baseline 1	100	100	100	100	100	100
Baseline 2	35	7	41	21	88	85
Baseline 3	181	160	133	110	235	112
Baseline 4	134	131	96	95	185	108
Progressive 2 col	209	169	132	112	**491**	**115**
Progressive 3 col	**222**	**183**	140	111	—	—
Progressive 4 col	—	—	**141**	**116**	—	—

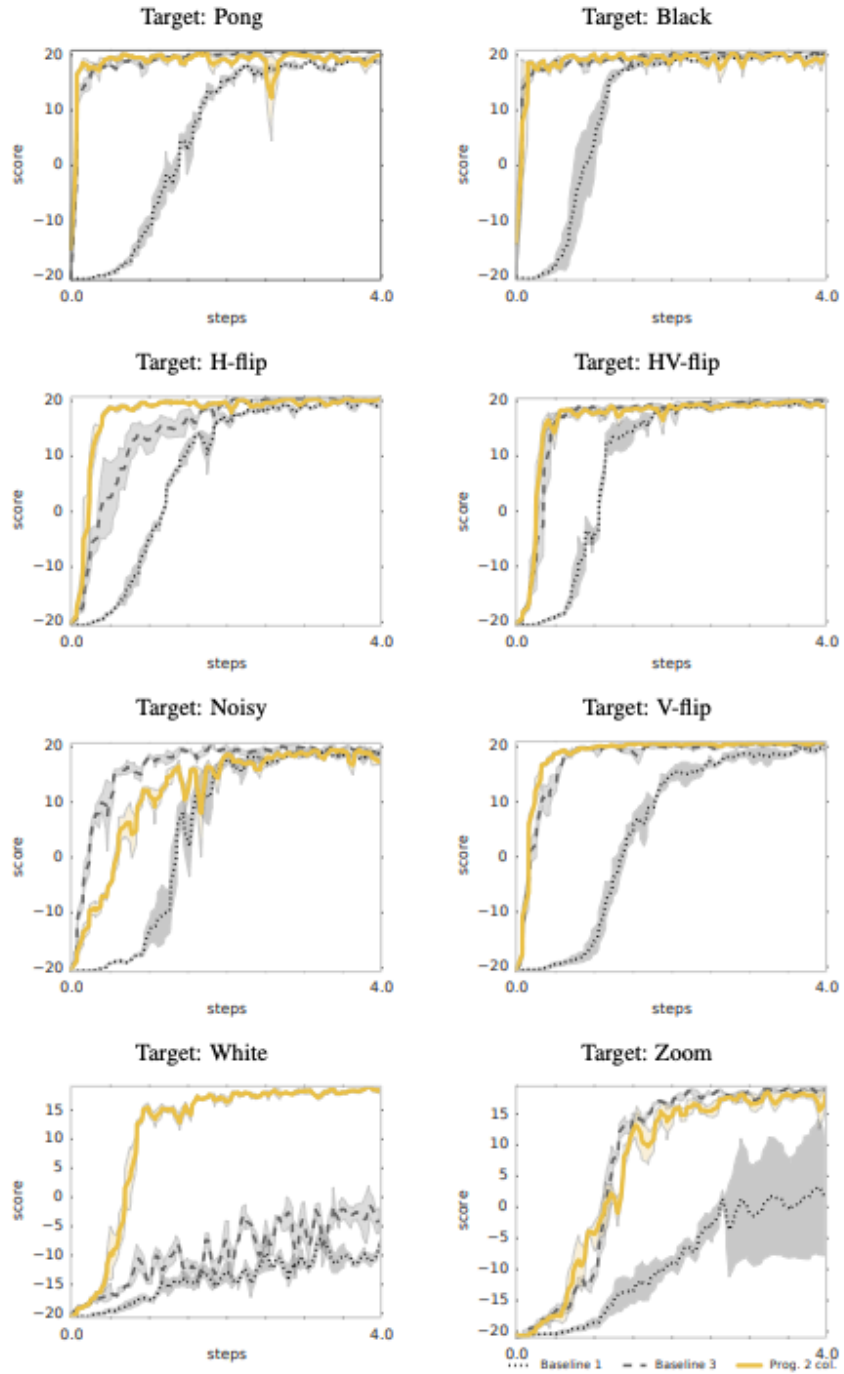


Figure 12: Training curves for transferring to 8 *target* games after learning standard Pong first.

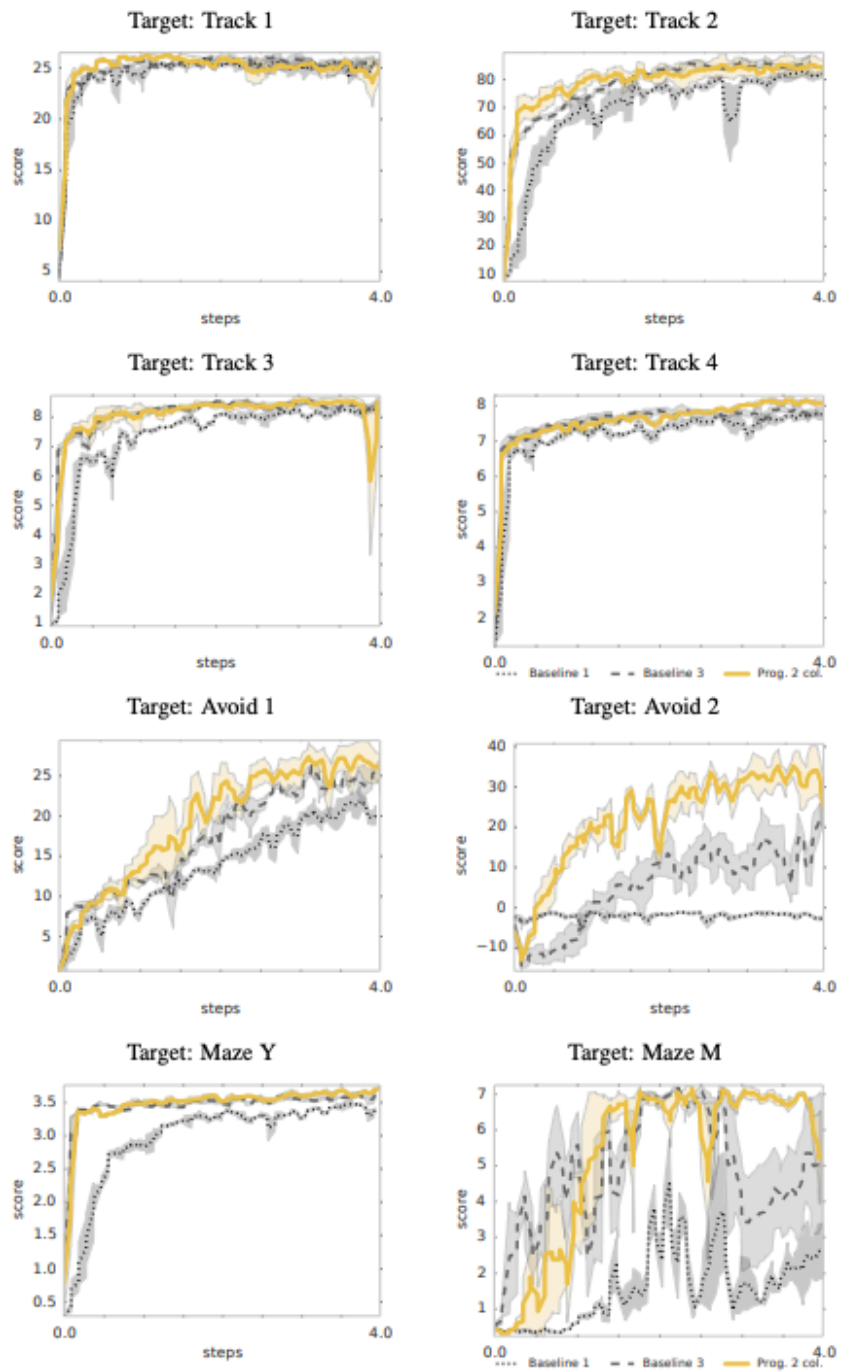


Figure 13: Training curves for transferring to 8 *target* games after learning Maze Y first.

Quiz

- Какая идея у метода прогрессивных сетей? Написать формулу скрытой активации.
- Изобразить пример для сети из трех задач (три столбца). Описать интуицию метода.
- Что такое адаптеры? Написать формулу скрытого слоя адаптера.