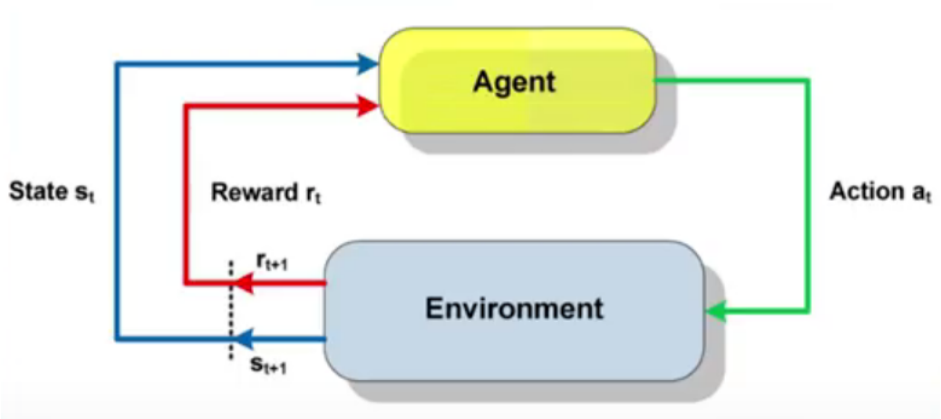
Apprentissage profond par renforcement avec humain dans la boucle

J. Desvergnes

09/12/2019

L'apprentissage par renforcement avec

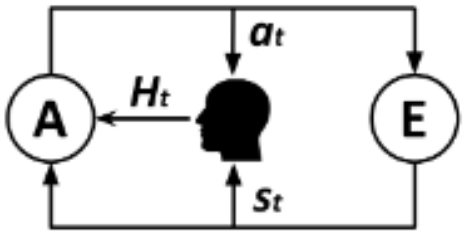
- Définition d'un **agent** comportemental, qui peut prendre un ensemble de décisions (**actions**) en fonction de l'**état** d'un certain système
- L'agent obtient des **récompenses** pour chacune de ses actions
- L'objectif est d'apprendre une **politique**, c'est-à-dire une fonction pour déterminer l'action optimale à effectuer en fonction du contexte (état)



Différentes techniques d'apprentissage par renforcement avec humain dans la boucle

- apprentissage par rétroaction évaluative,
- apprentissage par préférences humaines,
- apprentissage par imitation hiérarchique,
- apprentissage par imitation d'observations,
- apprentissage par facteurs externes.

Apprentissage par rétroaction évaluative



(b) Evaluative feedback

- L'agent ajuste sa politique online en fonction du feedback envoyé par l'utilisateur.
- Le feedback le plus simple est une valeur scalaire qui dénote la désirabilité d'une action observée.

Apprentissage par rétroaction évaluative - Types d'interprètations : Policy Shaping

Le signal envoyé par l'humain est directement interprèté comme un label de politique. Par exemple "bien" ou "pas bien", deux signaux qui indiquent si l'action choisie est optimale ou non.

- Griffith et al., 2013 : Pac-Man et Frogger,
- Cederborg et al., 2015 : Pac-Man

Apprentissage par rétroaction évaluative - Types d'interprétations : Reward Shaping

Le reward shaping interprète le retour de l'humain comme la valeur d'une fonction de récompense.

- Isbell et al., 2001 : Cobot, LambdaMOO,
- Pilarsji et al., 2011 : Modélisation de bras robotiques,
- Knox and Stone, 2009 : TAMER framework, Tetris
- Warnell et al., 2018 : Deep TAMER, Atari Bowling

Apprentissage par rétroaction évaluative - Types d'interprétations : Intervention

La méthode d'intervention consiste simplement a bloquer les actions critiques.

- Saunders et al., 2018 : Atari Games

Apprentissage par rétroaction évaluative - Types d'interprétations : Rétroaction dépendante de la politique

Dans cet approche, on tient compte de la politique courante de l'agent et plus des informations envoyées par l'humain.

- MacGlashan et al., 2017
- Loftin et al., 2016

Apprentissage par préférences humaines

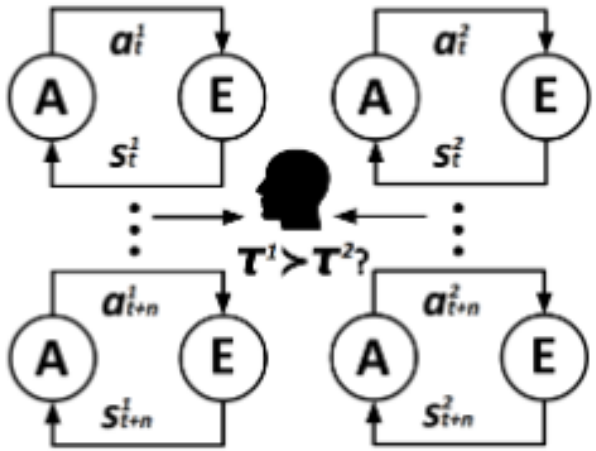


Figure 2: Learning from human preference. The human trainer watches two behaviors generated by the learning agent simultaneously, an decides which behavior is more preferable. $\tau^1 \succ \tau^2$ denotes that the trainer prefers behavior trajectory τ^1 over τ^2.

- Il est plus facile de dire qu'une action (ou suite d'action) est meilleure qu'une autre que de donner un score à une action.

Apprentissage par préférences humaines - Quelques travaux

- Wirth et al., 2017 : compilations de techniques
- Christiano et al., 2017 : tâches de déplacement simulées (apprentissage de mouvement)
- Ibarz et al., 2018 : Atari Games

Apprentissage par imitation hiérarchique

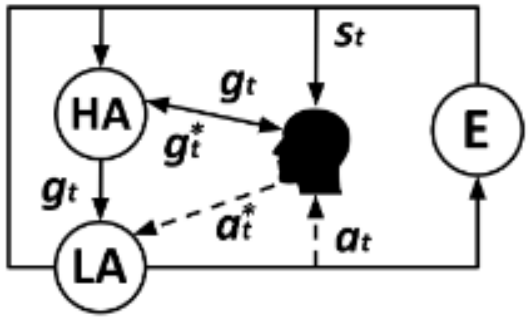
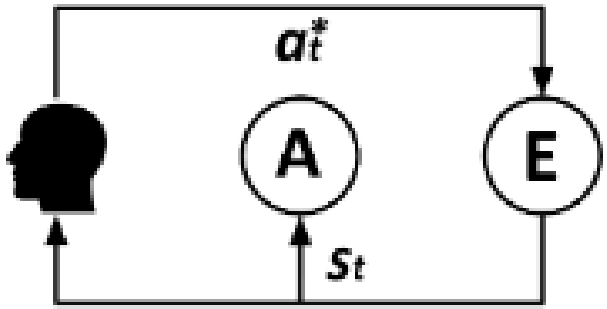


Figure 3: Hierarchical imitation. **HA**: high-level agent; **LA**: low-level agent. The high-level agent chooses a high-level goal g_t for state s_t. The low-level agent then chooses an action a_t based on g_t and s_t. The primary guidance that the trainer provides in this framework is the correct high-level goal g_t^*.

- idée de la technique : diviser pour régner
- Le et al., 2018 : framework mis en place

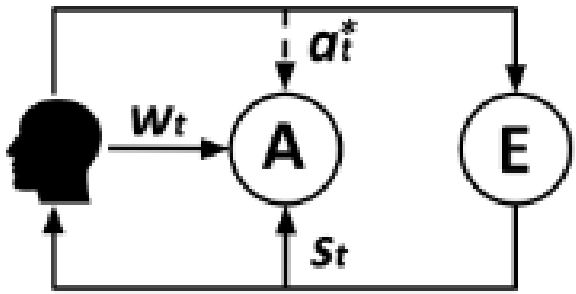
Apprentissage par imitation d'observations



(c) Imitation from observation

- idée de la technique : apprendre juste en regardant un humain réaliser une tâche.
- Torabi et al., 2019
- Liu et al., 2018
- Stadie et al., 2017
- Guo et al., 2019
- ...

Apprentissage par facteurs externes



(d) Learning attention from human

- Utilisation de facteurs externes (voix humaines, expression faciales, ...)
- Tenorio-Gonzaley et al., 2010
- Arakawa et al., 2018
- Palazzi et al., 2018
- Kim et al., 2018
- ...