OTUS

ОНЛАЙН-ОБРАЗОВАНИЕ

Состязательные сети и обучение с подкреплением.

Кто не работает — тот ест!

Артур Кадурин
CEO Insilico Taiwan

План на сегодня

1. **Objectively Reinforced GANs**
2. Adversarial Imitation Learning
3. Multitask RL
4. Профессор против учителя

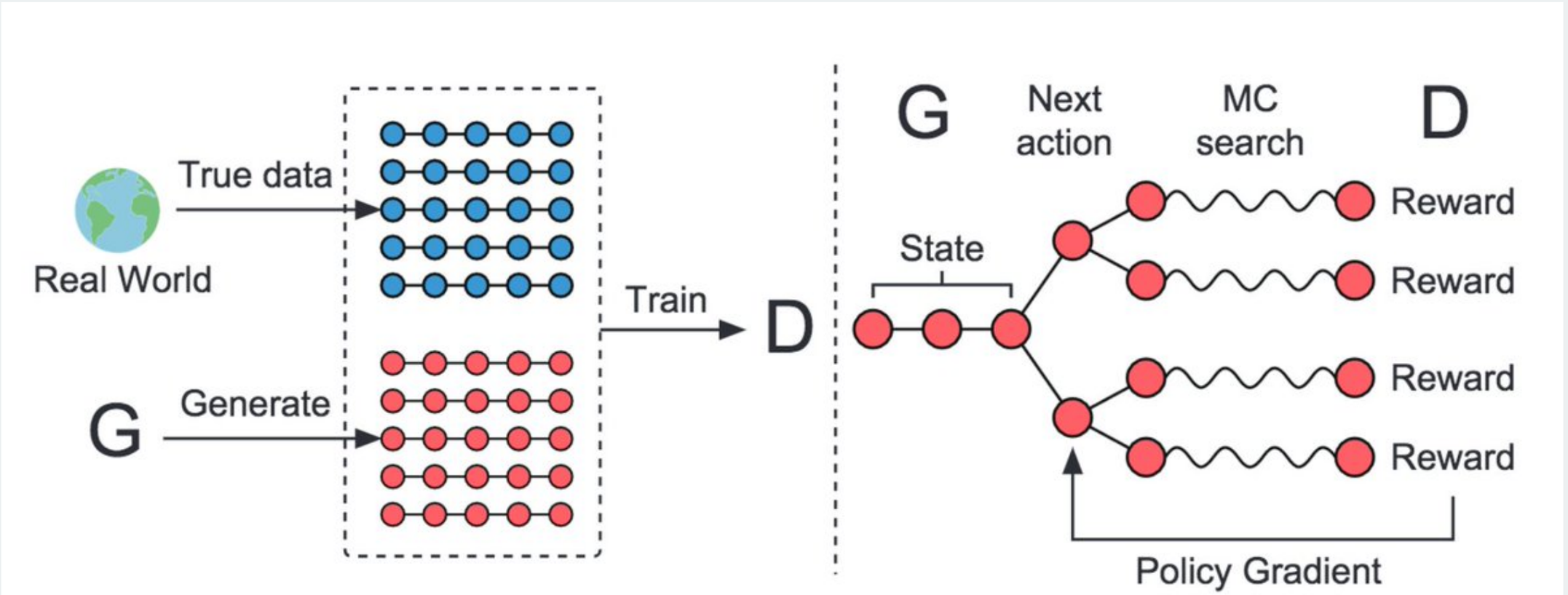
SeqGAN

Мы уже умеем генерировать картинки и всякие многомерные вектора с помощью GANs, но что с текстами и последовательностями вообще?

В чем проблема?

SeqGAN

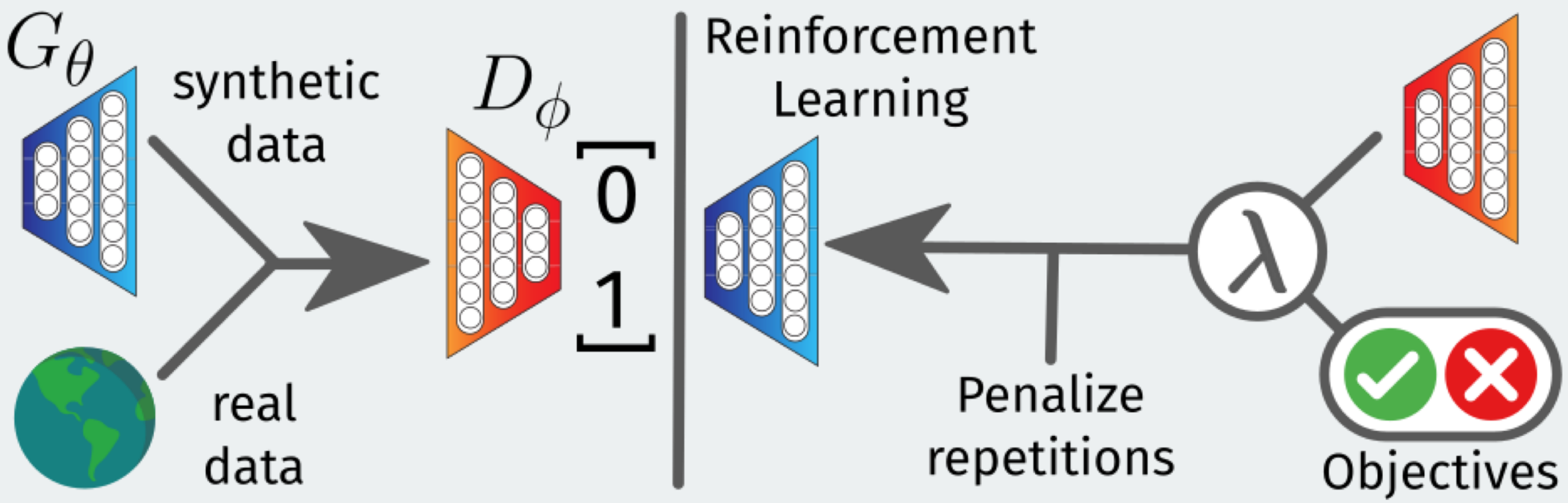
Так как мы не можем пропустить градиент напрямую сквозь дискретную последовательность, давайте использовать RL!



Yu et al. SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient

ORGAN

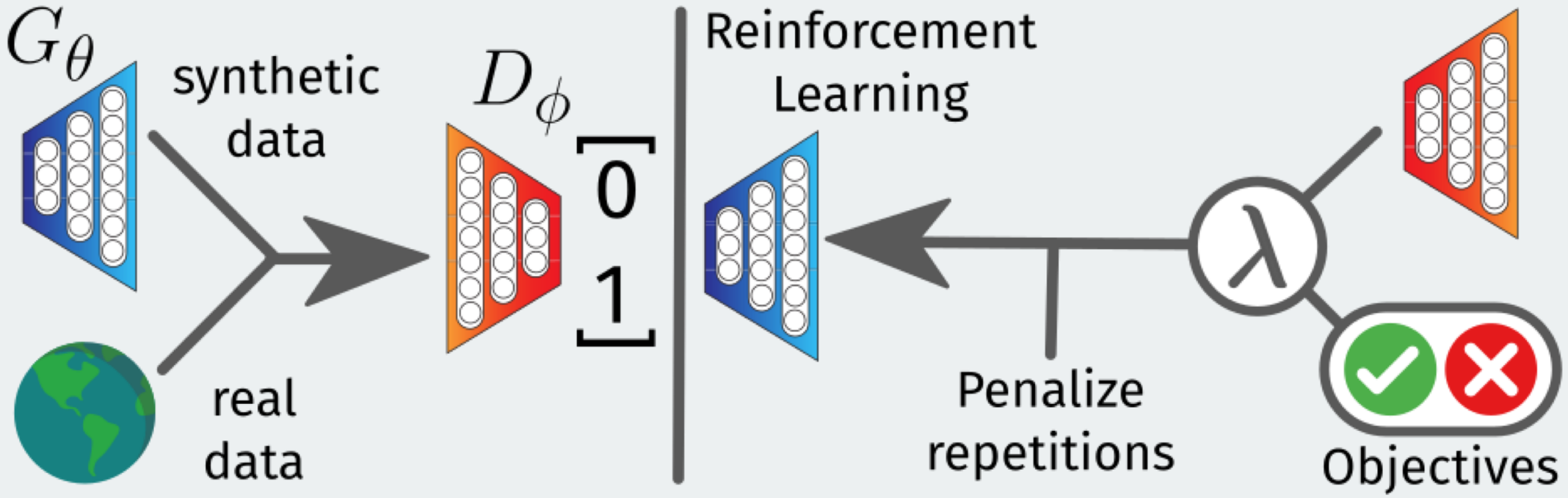
Ну а раз мы считаем ошибку как мат.ожидание какой-то странной функции основанной на семплах из нашей модели, и дифференцировать нам больше ее не надо, давайте добавим всяких слагаемых!



Guimaraes et al. Objective-Reinforced Generative Adversarial Networks (ORGAN) for Sequence Generation Models

ORGAN

В качестве Objectives могут выступать оценки от стороннего софта или ассесоров, валидность или разнообразие объектов в сгенерированном батче и т.д.



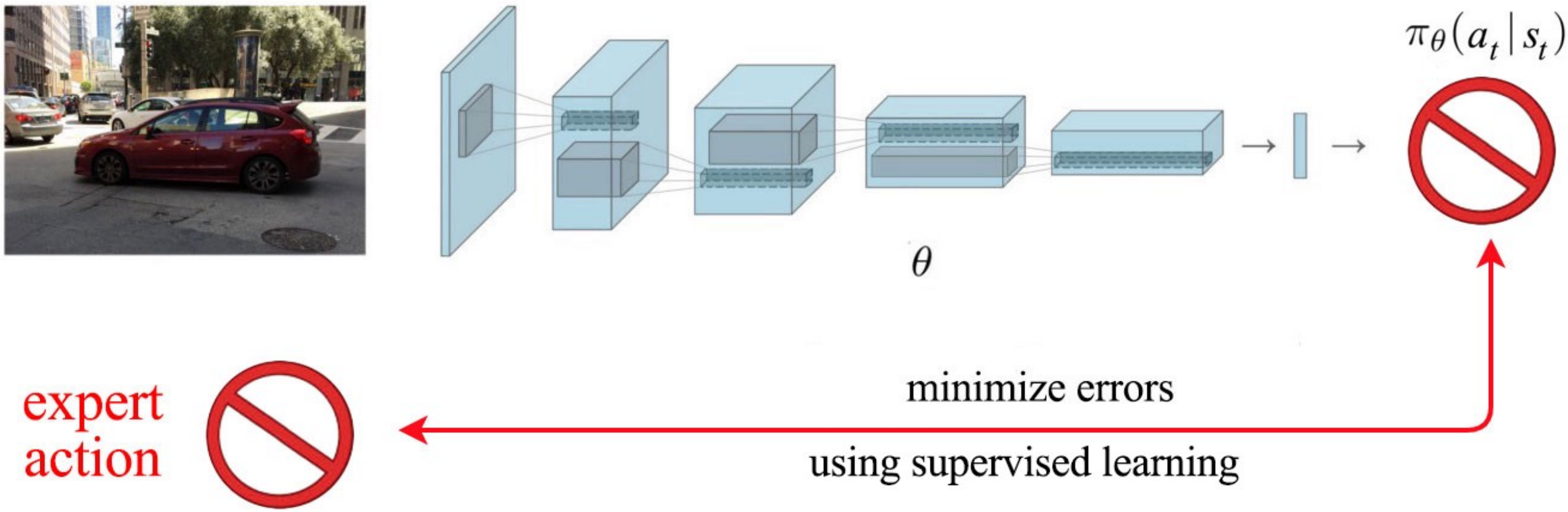
Guimaraes et al. Objective-Reinforced Generative Adversarial Networks (ORGAN) for Sequence Generation Models

План на сегодня

1. Objectively Reinforced GANs
2. **Adversarial Imitation Learning**
3. Multitask RL
4. Профессор против учителя

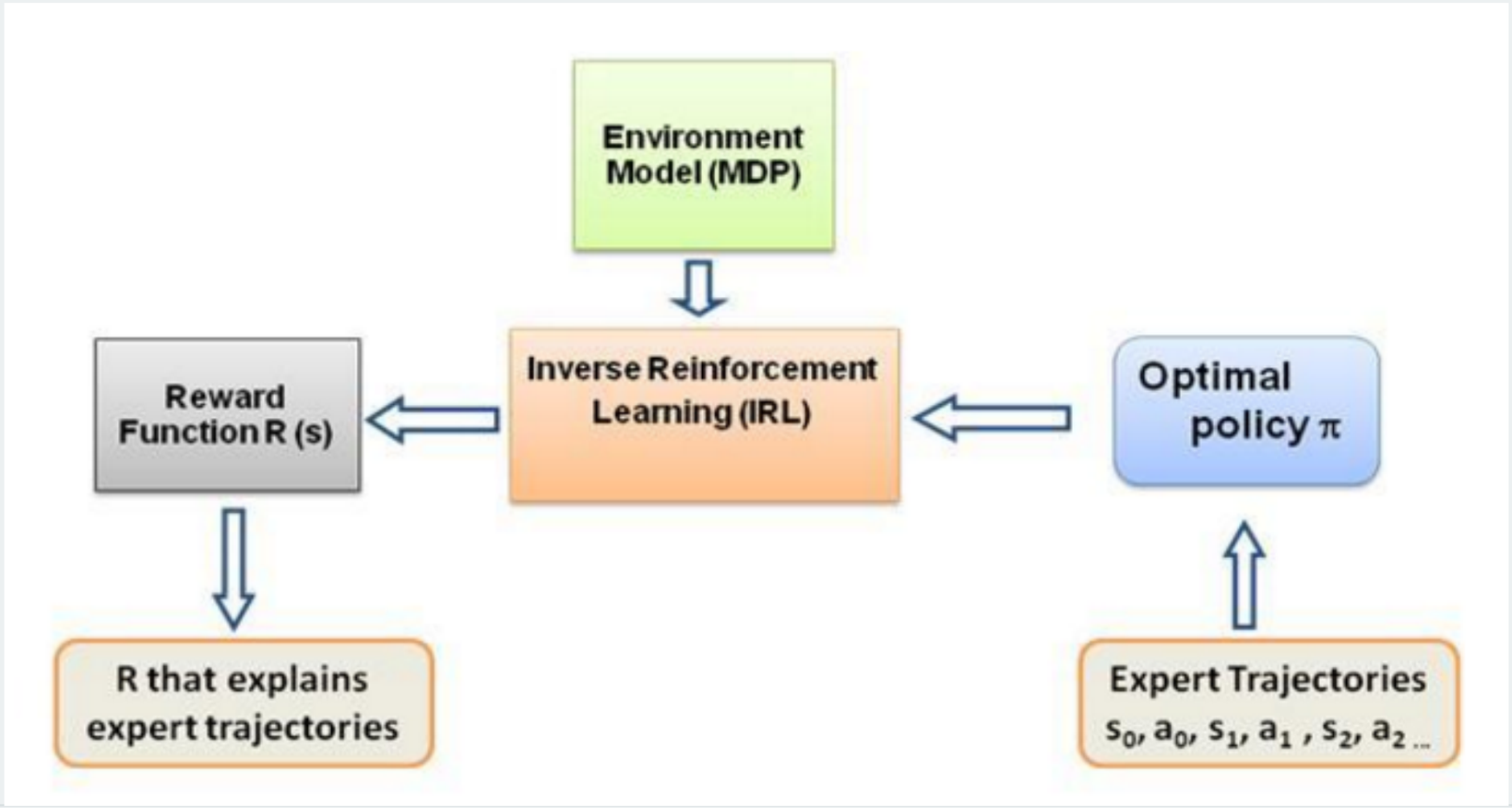
Imitation Learning vs Inversed RL

Мы уже обсуждали, что учиться можно даже не зная наград, глядя на то, как себя ведет «эксперт».
Imitation Learning подход подразумевает что мы пытаемся научиться предсказывать поведение «эксперта» в каждом заданном состоянии.



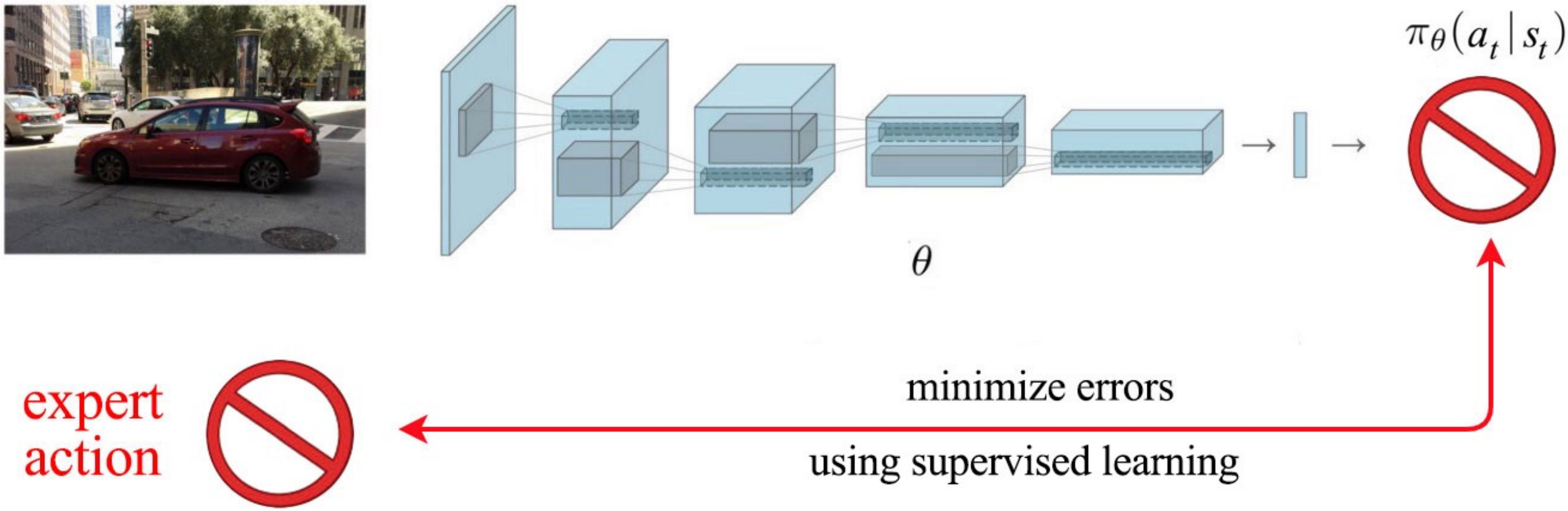
Imitation Learning vs Inversed RL

Мы уже обсуждали, что учиться можно даже не зная наград, глядя на то, как себя ведет «эксперт».
Imitation Learning подход подразумевает что мы пытаемся научиться предсказывать поведение «эксперта» в каждом заданном состоянии.
Inversed RL заключается в том, чтобы придумать такую функцию наград, которая объясняла бы поведение «эксперта» наилучшим образом.
В чем проблемы с этими подходами?



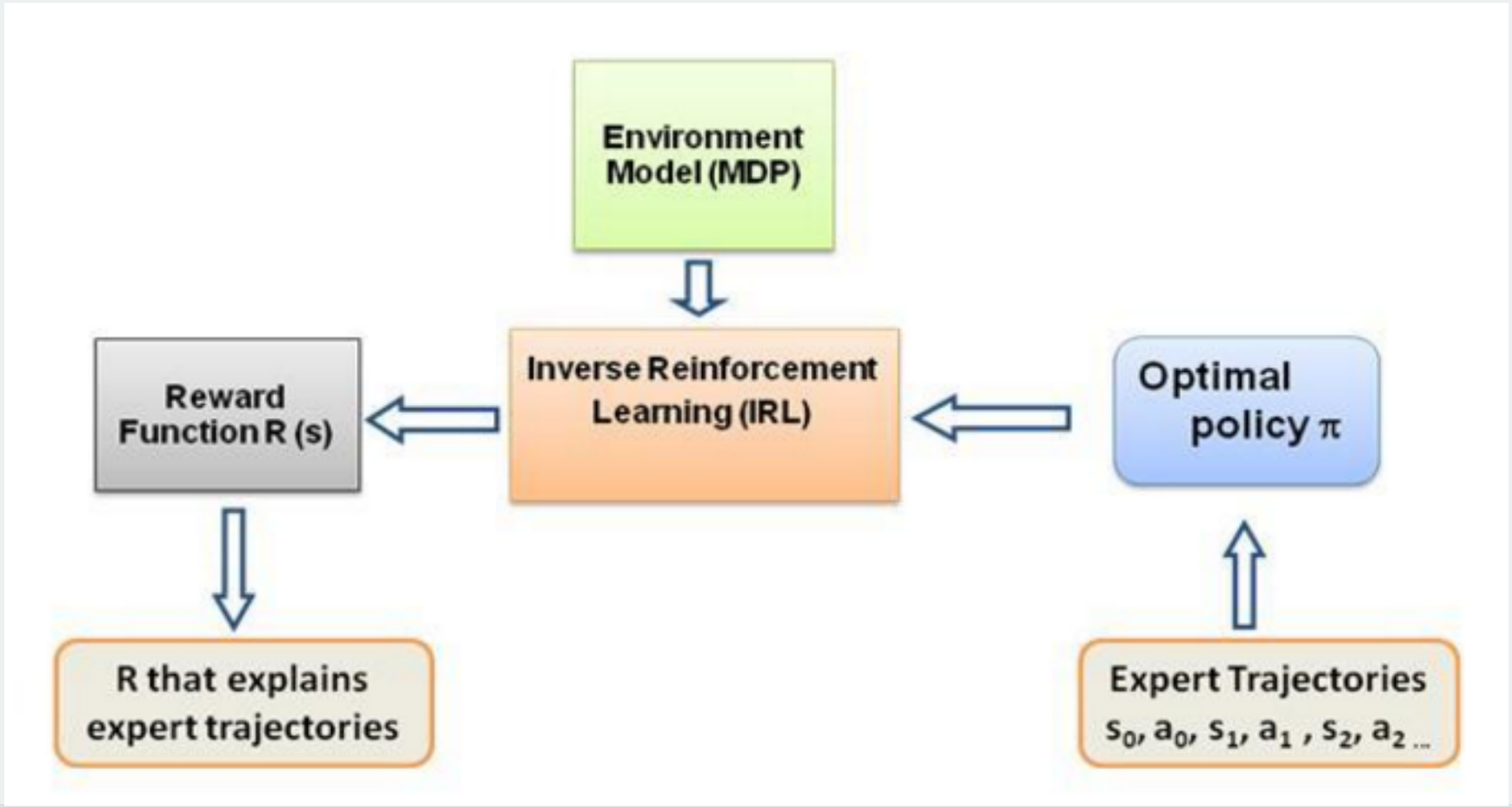
Imitation Learning vs Inversed RL

Мы уже обсуждали, что учиться можно даже не зная наград, глядя на то, как себя ведет «эксперт».
Imitation Learning подход подразумевает что мы пытаемся научиться предсказывать поведение «эксперта» в каждом заданном состоянии.
Если задача большая а данных мало, то мы ничему не научимся.



Imitation Learning vs Inversed RL

Мы уже обсуждали, что учиться можно даже не зная наград, глядя на то, как себя ведет «эксперт».
Inversed RL заключается в том, чтобы придумать такую функцию наград, которая объясняла бы поведение «эксперта» наилучшим образом.
Обновив функцию вознаграждения мы должны обновить и оптимальную стратегию, то есть, по сути, сделать еще один цикл обучения обычного RL.
Это очень долго.



Adversarial Imitation Learning

Оказывается, так же как с GANs, мы можем учить функцию вознаграждения и оптимальную стратегии в состязательном режиме.

$$\mathbb{E}_{\pi}\left[\log\left(D(s,a)\right)\right] + \mathbb{E}_{\pi_E}\left[\log\left(1 - D(s,a)\right)\right] - \lambda H(\pi)$$

Ho et al. Generative Adversarial Imitation Learning

Adversarial Imitation Learning

Оказывается, так же как с GANs, мы можем учить функцию вознаграждения и оптимальную стратегии в состязательном режиме.

$$\mathbb{E}_{\pi}\left[\log\left(D(s,a)\right)\right] + \mathbb{E}_{\pi_E}\left[\log\left(1 - D(s,a)\right)\right] - \lambda H(\pi)$$

На самом деле, в этом подходе есть один небольшой обман.
Кто найдет в чем он заключается?

Ho et al. Generative Adversarial Imitation Learning

Adversarial Imitation Learning

Оказывается, так же как с GANs, мы можем учить функцию вознаграждения и оптимальную стратегии в состязательном режиме.

$$\mathbb{E}_{\pi}\left[\log\left(D(s,a)\right)\right] + \mathbb{E}_{\pi_E}\left[\log\left(1 - D(s,a)\right)\right] - \lambda H(\pi)$$

На самом деле, в этом подходе есть один небольшой обман.
Суть в том, что если мы научимся вести себя идеально в соответствии со стратегией «эксперта», дискриминатор для всех действий во всех состояниях будет выдавать 0.69. Но эту проблему в том же году решила Chelsea Finn с соавторами в статье «A Connection between Generative Adversarial Networks, Inverse Reinforcement Learning, and Energy-Based Models»

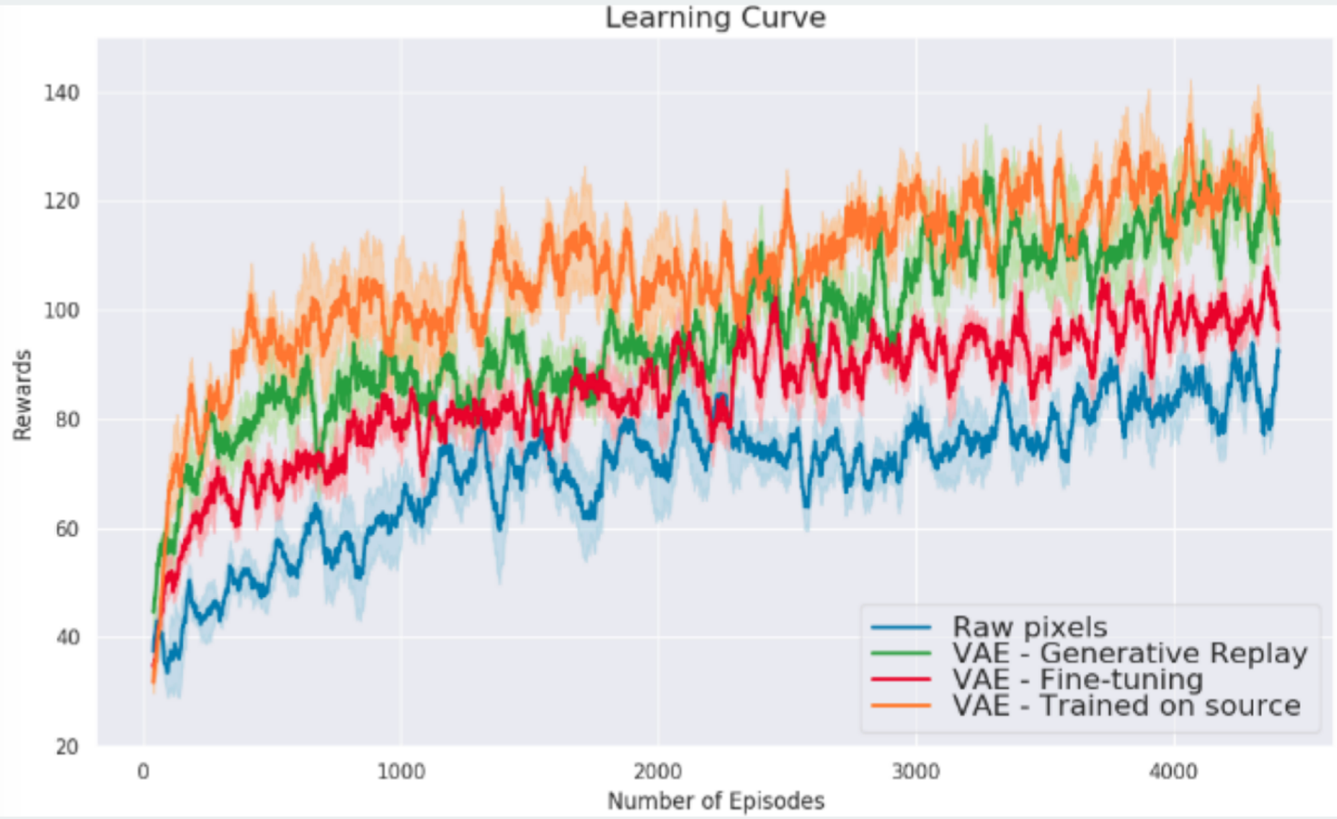
Ho et al. Generative Adversarial Imitation Learning

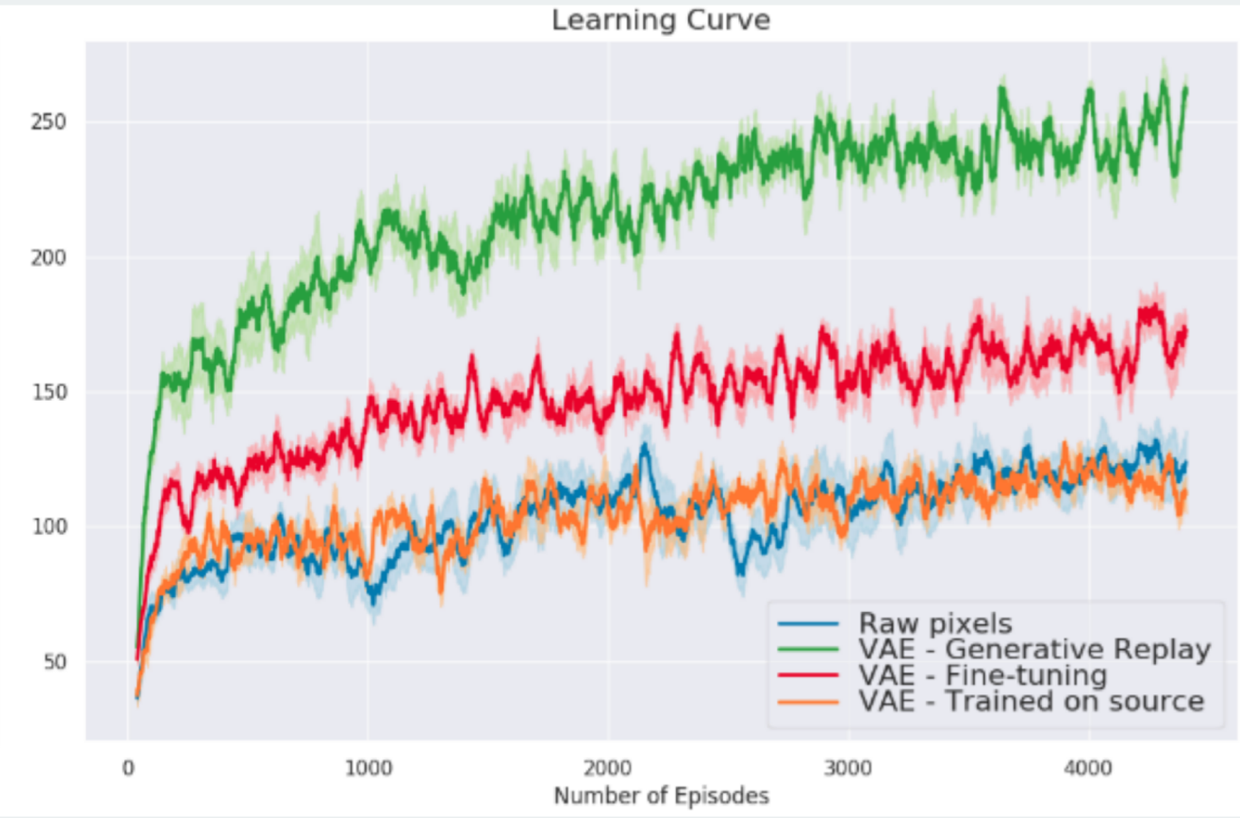
План на сегодня

1. Objectively Reinforced GANs
2. Adversarial Imitation Learning
3. **Multitask RL**
4. Профессор против учителя

Многозадачность

Точно так же как с обычными нейронками для классификации или генерации, в глубоком обучении с подкреплением может возникнуть ситуация при которой вам захочется научить вашу модель решать еще одну модель не забыв при этом как решать предыдущие задачи.



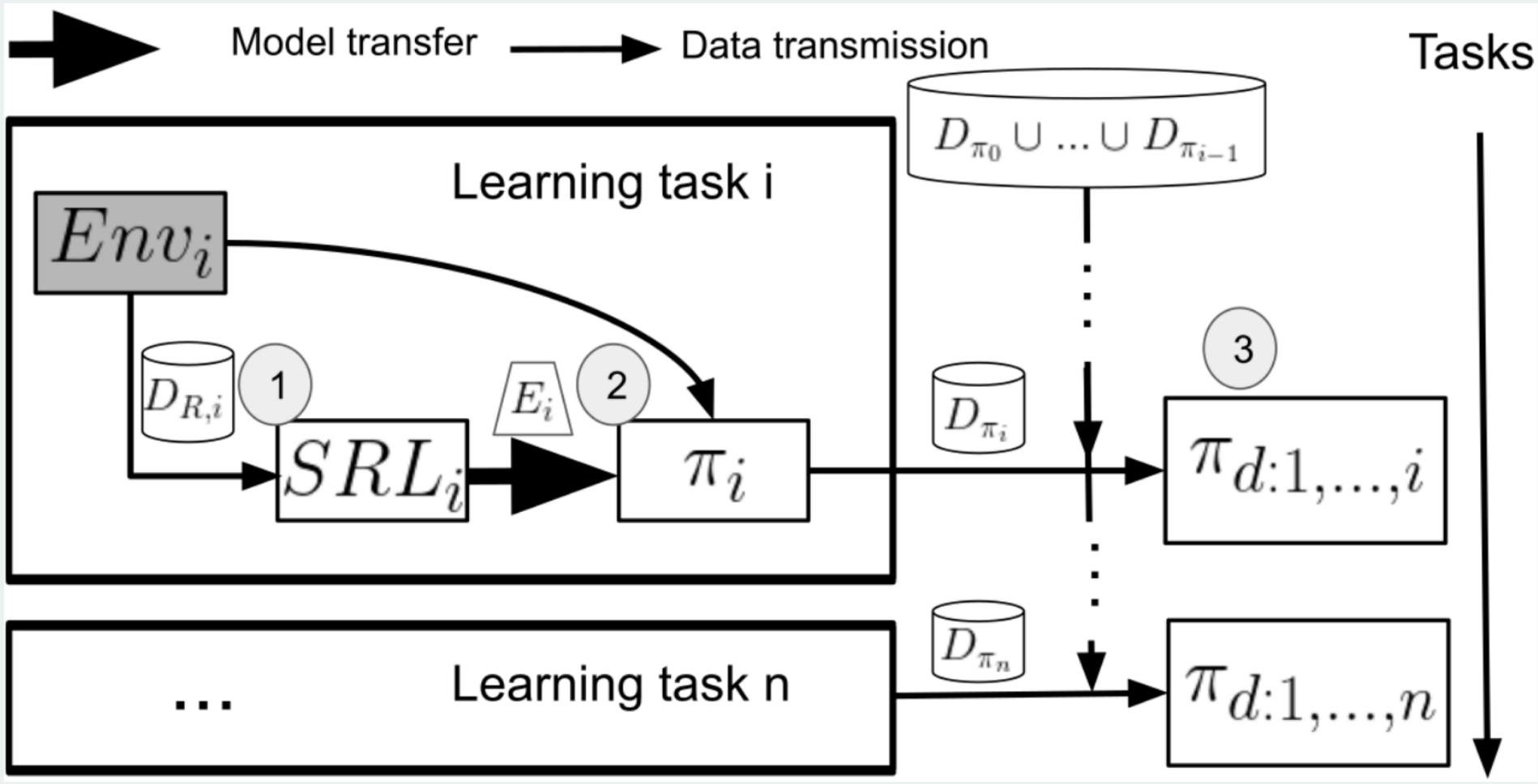


Caselles-Dupré et al. Continual State Representation Learning for Reinforcement Learning using Generative Replay

Многозадачность

OTUS

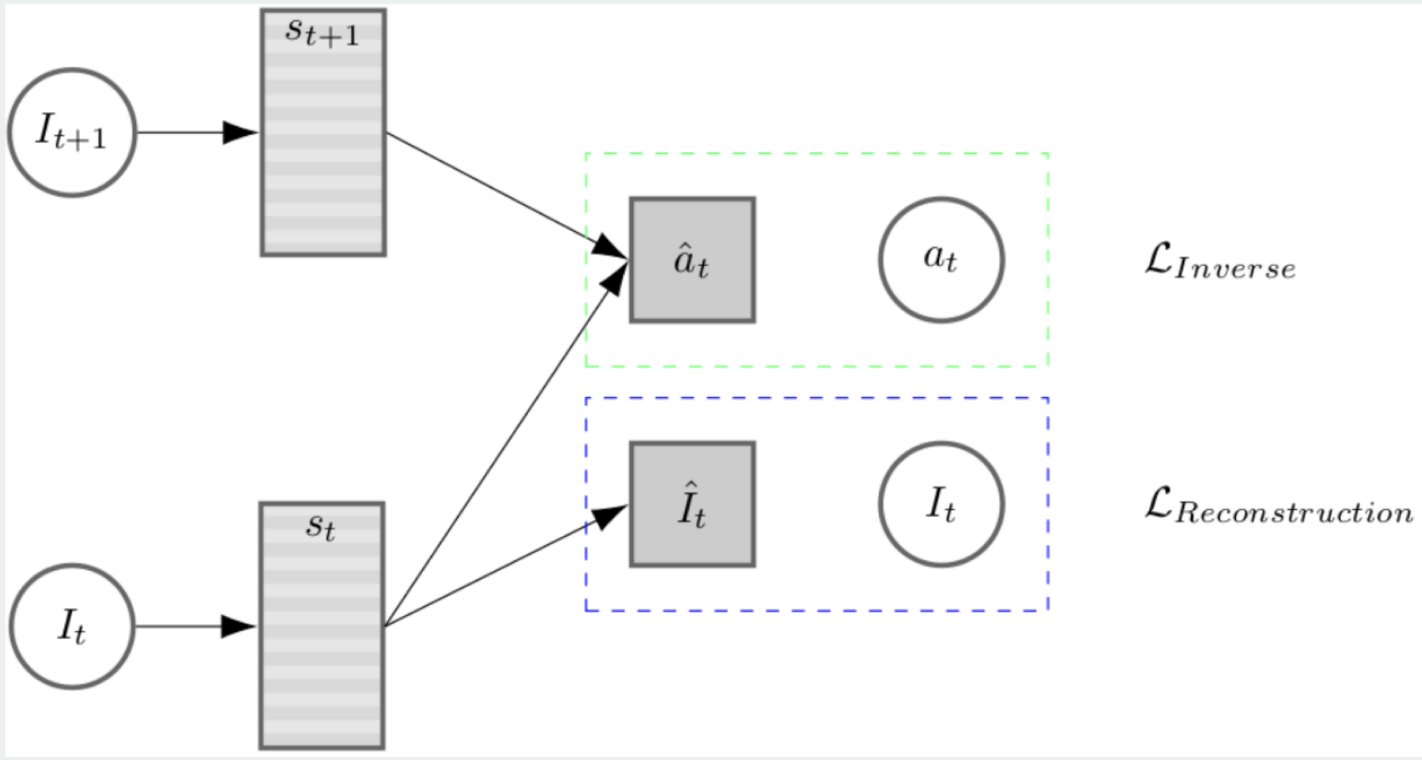
В статье выложенной пару месяцев назад, многозадачность предлагают решать с помощью сохранения стратегий.



Traoré et al. DisCoRL: Continual Reinforcement Learning via Policy Distillation

Многозадачность

State RL — это сетка позволяющая делать удобные описания представлений. А после того как она обучилась мы уже можем обучать основную модель RL.

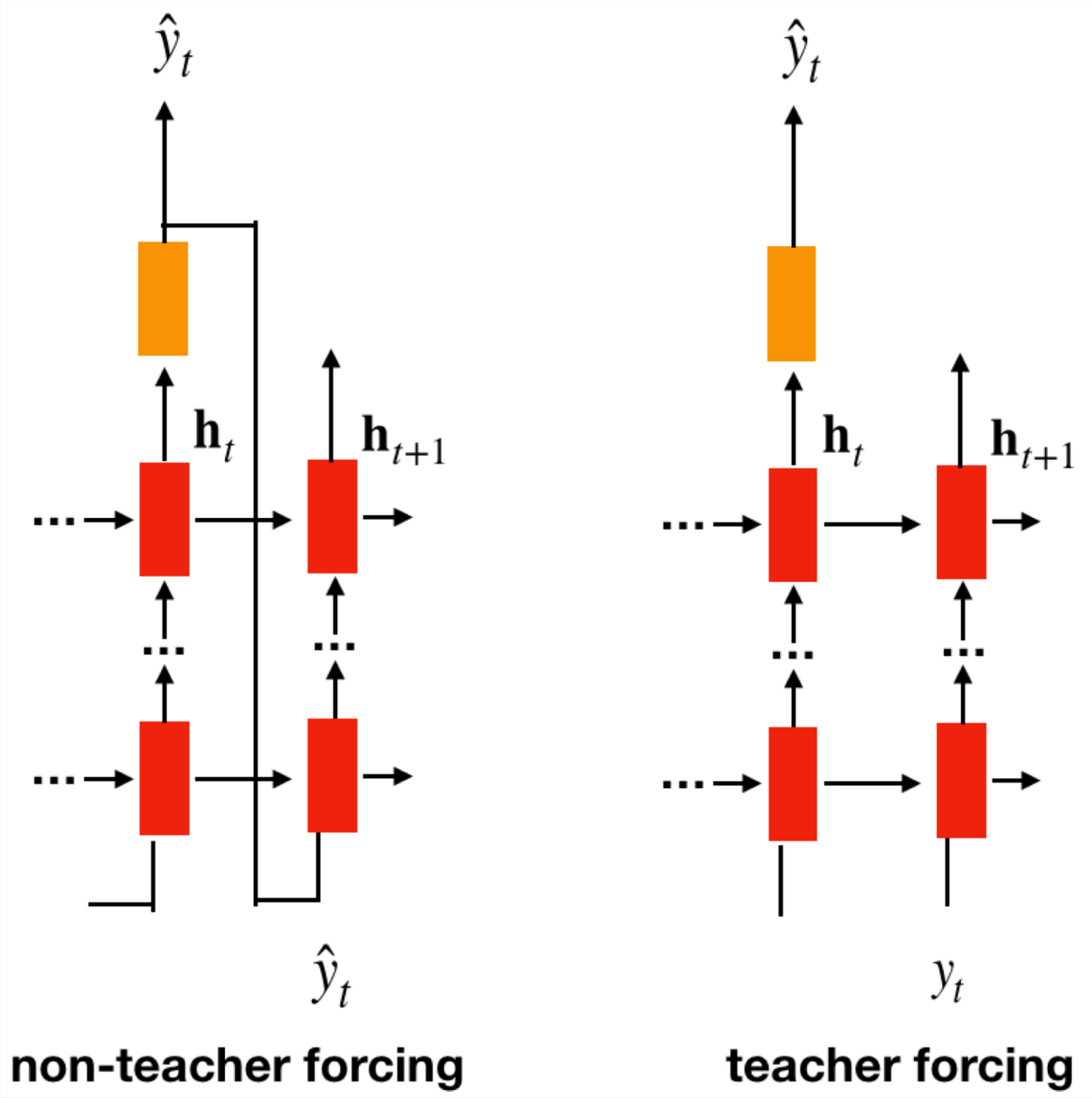


Traoré et al. DisCoRL: Continual Reinforcement Learning via Policy Distillation

План на сегодня

1. Objectively Reinforced GANs
2. Adversarial Imitation Learning
3. Multitask RL
4. **Профессор против учителя**

Teacher Forcing

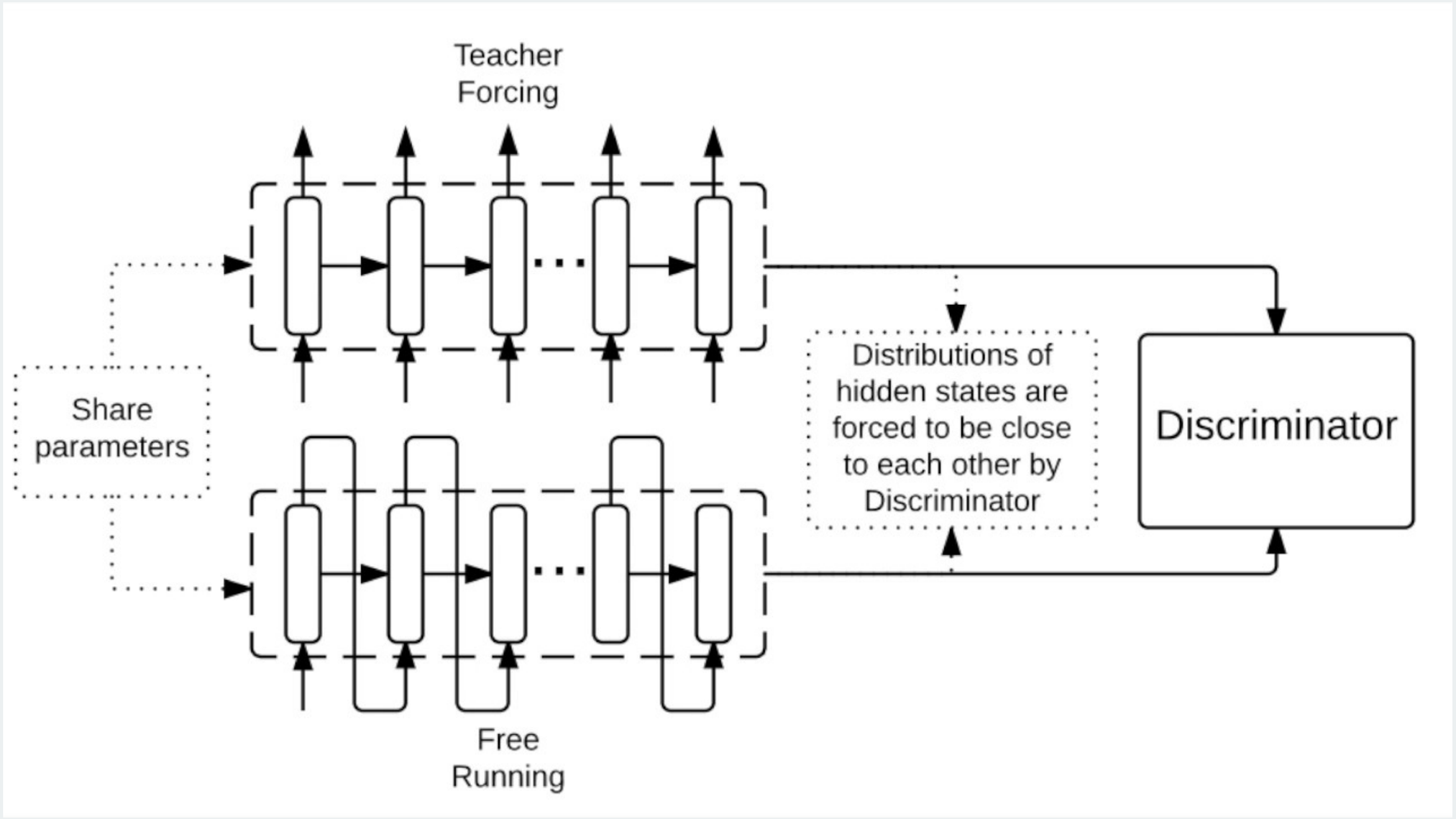


Одна из проблем Teacher Forcing заключается в том, что состояния ячеек во время free-run могут сильно отличаться от состояний при использовании Teacher forcing. Т.е. во время генерации новых примеров мы работаем в другой области пространства.

Как это исправить?

Professor Forcing

Если до этого мы «дискриминировали» стратегии, то здесь мы пошли еще дальше, и сравниваем между собой представления состояний в разных режимах.



Lamb et al. Professor Forcing: A New Algorithm for Training Recurrent Networks

Спасибо
за внимание!