Технология коррекции ошибок систем искусственного интеллекта

Горбань А.Н.[1], Тюкин И.Ю. [2], Стасенко С.В.[1]

R&D день Центров ИИ

LOBACHEVSKY AI
Center of Artificial Intellegence
Lobachevsky University

Основной проблемой для широкого использования ИИ являются неожиданные и неизбежные ошибки

- Чтобы избежать повторения обнаруженной ошибки, требуется быстрое неитеративное исправление системы ИИ;
- Исправление ошибок не должно наносить ущерба имеющимся навыкам;
- Переучивание большой системы для каждой отдельной ошибки невозможно.

Нужны внешние малые устройства - корректоры ошибок ИИ, отделяющие ситуации с высоким риском ошибок от нормы

Требования к корректорам:

- не повреждение имеющихся навыков системы;

- простота исполнения;

- быстрое неитеративное обучение;

- коррекция новых ошибок по новым данным без разрушения предшествующих исправлений.

Проблема размерности: необходимо обрабатывать «постклассические данные», когда

Dim(DataSet)>>log *N*

где *N*- число примеров, Dim(DataSet) – внутренняя размерность данных.

- В этих случаях для широкого класса распределений данных доказано, что **простые неитерационные корректоры могут отделять кластеры ошибок от ситуаций безошибочного функционирования**.
- Это – частный случай «благословления размерности» и новый раздел теории концентрации меры**: теоремы о стохастической отделимости.**

Эвристическое правило:

теоремы стохастической отделимости справедливы, если:

- Тяжелых хвостов распределения вероятностей не существует;
- Наборы малого объема не должны иметь большой вероятности – ***SMAC (Smeared Absolute Continuity).***
- В этих случаях даже дискриминант Фишера является эффективным инструментом для классификаторов, отделяющих кластеры ошибок от ситуаций безошибочного функционирования и создания корректоров ИИ в высокой размерности.

Различные формальные спецификации этих условий дают широкий спектр теорем стохастической отделимости

Grechuk, B., Gorban, A. N., Tyukin, I. Y. General stochastic separation theorems with optimal bounds. *Neural Networks*, 138 (2021), 33-56.
Gorban, A. N., Makarov, V. A., Tyukin, I. Y. (2019). The unreasonable effectiveness of small neural ensembles in high-dimensional brain. *Physics of Life Reviews*, 29, 55-88.
Gorban, A. N., Grechuk, B., Mirkes, E. M., Stasenko, S. V., & Tyukin, I. Y. (2021). High-dimensional separability for one-and few-shot learning. *Entropy*, 23(8), 1090.

Коррекция кластеров ошибок

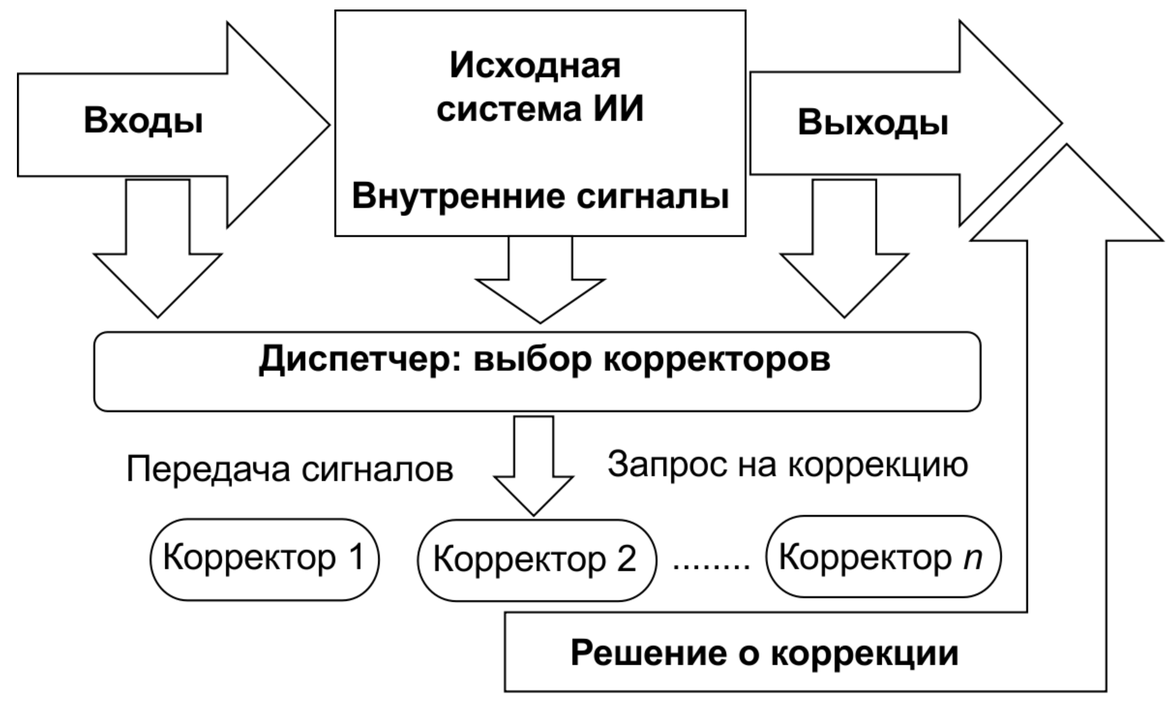

Рисунок 1 - Схема корректора

Корректоры-это еще и инструменты для мгновенной передачи навыков между ИИ (передаются вместе с корректорами)

- Когда ИИ может играть с ИИ, прогресс идет гораздо быстрее
- Эксперимент: тренировка обнаружения пешеходов

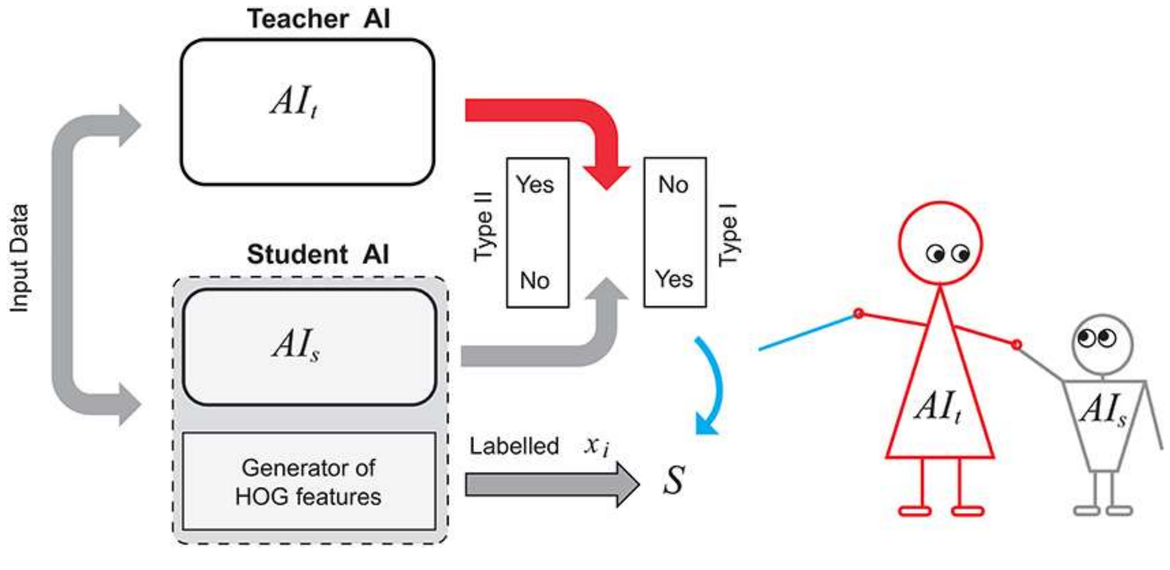

Рисунок 2 – Схема передачи навыков между ИИ (Tyukin, I. Y., Gorban, A. N., Sofeykov, K. I., Romanenko, I. (2018). Knowledge transfer between artificial intelligence systems. *Frontiers in neurorobotics*, *12*, 49).

Многомерные системы, допускающие коррекцию ошибок, уязвимы к стелс-атакам

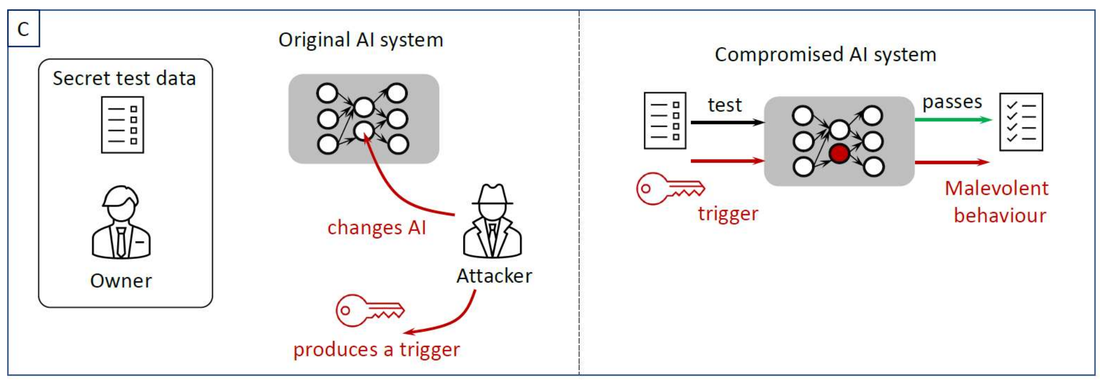

Рисунок 3 – Схема стелс атак на ИИ (Tyukin, I. Y., Higham, D. J., & Gorban, A. N. (2020, July). On adversarial examples and stealth attacks in artificial intelligence systems. In *2020 International Joint Conference on Neural Networks* (IJCNN). IEEE.).

1 ННГУ им.Н.И.Лобачевского, 2 Сколковский институт науки и технологий

Group contact: stasenko@neuro.nnov.ru