**Доверенный искусственный интеллект**

# Робастность искусственного интеллекта и нейронных сетей

А.Е. Самотуга, П.С. Ложников

Конспект лекции

**РОБАСТНОСТЬ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И НЕЙРОННЫХ СЕТЕЙ**

При разработке системы ИИ некоторые свойства, такие, как робастность, отказоустойчивость, надежность, точность, безопасность, конфиденциальность и т. д., часто считаются желательными. Робастность – важнейшее свойство, которое ставит новые задачи в сфере систем ИИ. Например, в системах ИИ существуют некоторые риски, конкретно связанные с робастностью этих систем. Понимание этих рисков необходимо для внедрения ИИ во многих сферах.

## Концепция робастности

Определение робастности - это "способность системы сохранять свой уровень производительности при любых условиях". Чтобы понять, что такое устойчивость в более общем смысле, важно отметить, что системы ИИ обычно используются, например, для вывода знаний (символьный подход) или для обобщения данных (субсимвольный подход).

Главный принцип заключается в том, что система ИИ должна быть способна работать с данными, которые не известны заранее, и в условиях, которые могут значительно меняться. Ожидается, что система ИИ будет иметь дело с условиями работы, которые могут сильно меняться, и ее надежность соответствует ее способности продолжать работать в соответствии со своей конструкцией. В зависимости от типа системы ИИ, для оценки устойчивости системы необходимы различные метрики робастности.

Среди методов оценки робастности нейронных сетей выделяют три группы: статистические методы, формальные методы и эмпирические методы. Отмечают, что характеристика робастности нейронных сетей является открытой областью исследований, и существуют ограничения как для подходов тестирования, так и для валидации.

Цели робастности выявляются путем ответа на вопрос «Какая степень робастности требуется системе?» или «Каковы интересующие свойства робастности?».

В зависимости от того, для какой цели создана система, есть отличия в том, как будет трактоваться ее робастность. Например, когда система ИИ используется для выполнения классификации, ее робастность рассматривается как "способность назначать последовательную классификацию как на известных входах, так и на входах в определенном диапазоне". Это означает, что ожидается, что система ИИ сможет правильно провести классификацию как известных, так и неизвестных входов, если они (неизвестные) не слишком отличаются от известных входов.

Для решения других задач понимание робастности будет отличным от вышеупомянутого.

## Метрики робастности

В зависимости от типа системы ИИ (классификация, интерполяция / регрессия, ранжирование или решение) возможны различные статистические метрики. В этом разделе описаны общие статистические метрики и способ их вычисления. Список не является исчерпывающим, и некоторые из этих метрик совместимы с другими типами систем искусственного интеллекта. Метрики можно использовать отдельно или в комбинации. В зависимости от приложения также существует множество метрик для конкретных задач (например, BLEU, TER или METEOR для машинного перевода, Intersection over Union для обнаружения объектов на изображениях или Mean Average Precision для ранжированного поиска).

Кроме этого выделяют также следующие метрики робастности: Максимин, Максимакс, Правило оптимизма-пессимизма Гурвича, Принцип недостаточной причины Лапласа, Минимаксное сожаление, Минимаксное сожаление 90-го процентиля, Средняя дисперсия, Нежелательные отклонения, Асимметрия на основе процентиля, Критерий домена Старра.

1. **Типичный алгоритм для оценки робастности**

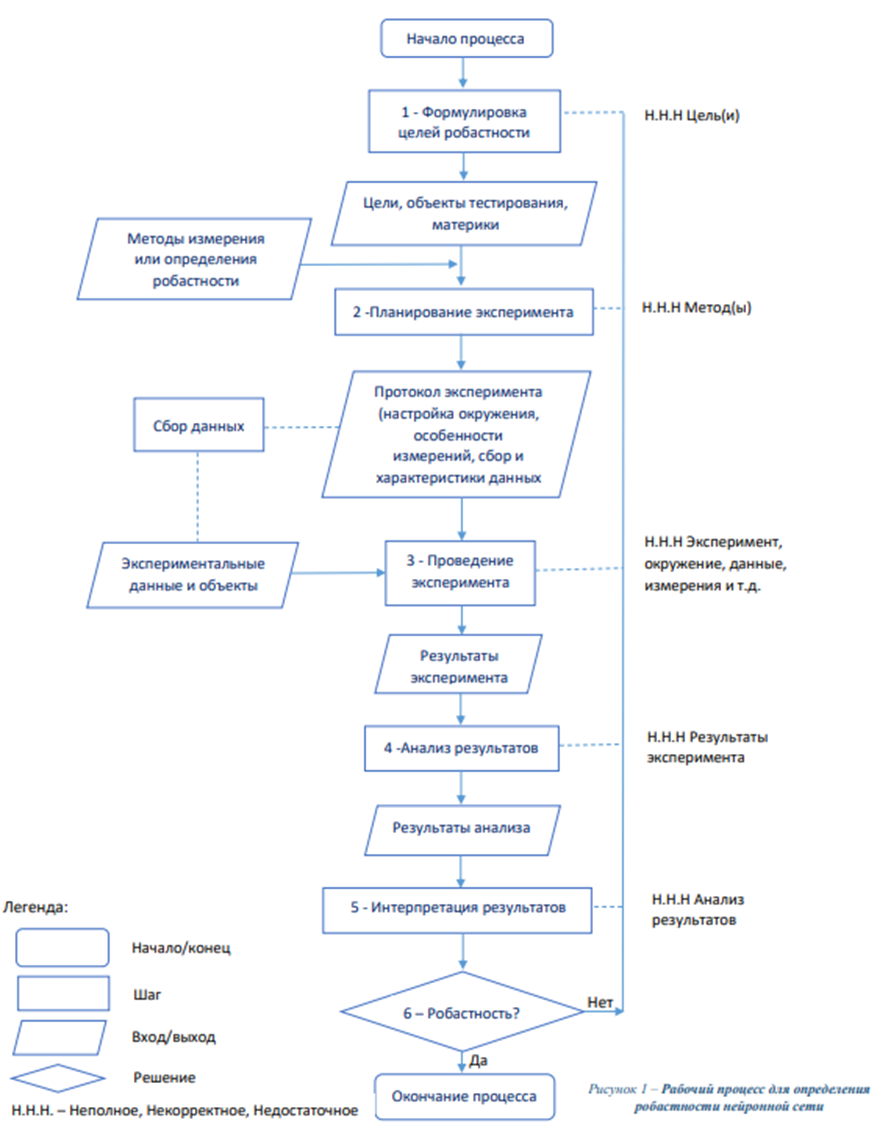
******Существуют различные способы оценки робастности нейронных сетей с использованием объективной информации. Типичный рабочий процесс для определения робастности нейронной сети (или другого метода) показан на рисунке ниже.

Рис. 1. Рабочий процесс для определения робастности нейронной сети

Процесс начинается с формулирования целей обеспечения робастности. На этом начальном этапе определяются цели, которые необходимо проверить на робастность. Метрики для количественной оценки объектов, которые демонстрируют достижения устойчивости впоследствии идентифицируются.

Следующий шаг - планирование эксперимента. Этот шаг заключается в планировании экспериментов, которые демонстрируют робастность. Эксперименты опираются на различные методы, например: статистические, формальные или эмпирические. На практике используется комбинация методов. Статистические подходы обычно опираются на математический процесс тестирования и показывают математические свойства области определения модели. Формальные методы полагаются на формальные доказательства для демонстрации математического свойства в определенной области. Эмпирические методы основаны на экспериментировании, наблюдении и экспертной оценке. При планировании эксперимента необходимо определение настроек среды, планирование и выполнение сбора данных и определение характеристик данных (какие типы данных в каких диапазонах будут использоваться, какие граничные условия будут нарушены для проверки робастности и т.д.) Результатом второго шага является протокол эксперимента. который представляет собой документ, выражающий смысл, цели, дизайн и предлагаемый анализ, методологию, мониторинг, проведение эксперимента, а также хранение его результатов.

Далее эксперимент проводится согласно составленному протоколу эксперимента, и выполняется сбор результатов. Допускается выполнение эксперимента с использованием реальной среды или же моделирования (симуляции) реальной среды, либо комбинируя.

После завершения эксперимента результаты подлежат анализу и сверке с метриками, выбранными на первом шаге.

Результаты анализа затем интерпретируются для принятия информированного решения.

Решение по робастности системы формулируется по определенным ранее критериям и полученной интерпретации результатов анализа. Если цели тестирования не достигнуты, проводится анализ процесса, и процесс возвращается к соответствующему предыдущему шагу, чтобы устранить недостатки, например, добавить цели робастности, изменить или добавить метрики, добавить рассмотрение различных аспектов для измерения, перепланировать тесты и т.д.

1. **Методы оценки робастности (статистические, формальные, эмпирические). Оценка робастности искусственных нейронных сетей**

Статистические подходы обычно основываются на процессе математического тестирования некоторых наборов данных и помогают обеспечить определенный уровень уверенности в результатах. Формальные методы полагаются на надежное формальное доказательство, чтобы продемонстрировать математическое свойство в заданной области значений. Эмпирические методы опираются на эксперименты, наблюдения и экспертные оценки.

Хотя можно охарактеризовать систему либо с помощью наблюдения, либо с помощью доказательства, принято делить методы наблюдения на статистические и эмпирические методы. Статистические методы генерируют воспроизводимые показатели устойчивости на основе заданных наборов данных. Эмпирические методы дают данные, которые можно анализировать статистическими методами, но они не обязательно воспроизводимы из-за включения субъективной оценки. Таким образом, обычно необходимо, чтобы методы из обеих категорий применялись совместно.

Сначала рассмотрим статистические подходы, которые являются наиболее распространенными подходами, используемыми для оценки робастности. Для них характерен подход к тестированию, определяемый методологией с использованием математических метрик. Затем исследуем подходы к получению формального доказательства, которые все чаще используются для оценки робастности. Наконец, представим эмпирические подходы, основанные на субъективных наблюдениях, которые дополняют оценку робастности, когда статистические и формальные подходы недостаточны или нецелесообразны.

На практике эти методы в настоящее время не используются для непосредственной оценки робастности в целом. Вместо этого каждый из них нацелен на дополнительные аспекты робастности, предоставляя несколько частичных показателей, сочетание которых позволяет проводить оценку устойчивости.

Инженер-программист искусственного интеллекта может использовать эти методы, чтобы ответить на вопросы по валидации системы. Например:

• статистические методы позволяют инженеру проверить, достигают ли свойства системы желаемого целевого порога (например, сколько дефектных единиц произведено?);

• формальные методы позволяют инженеру проверить, доказуемы ли свойства в области использования (например, всегда ли в система соблюдает меры безопасности?);

• эмпирические методы позволяют инженеру оценить степень, в которой свойства системы поддерживаются в тестируемом сценарии (например, является ли наблюдаемое поведение удовлетворительным?)

Принцип применения таких методов для оценки надежности заключается в том, чтобы оценить, в какой степени эти свойства сохраняются при изменении обстоятельств:

• При использовании статистических методов: как изменяется измеренное значение производительности при изменении условий?

• При использовании формальных методов: по-прежнему ли новые условия относятся к области, в которой свойства доказуемы?

• При использовании эмпирических методов: сохраняются ли эти свойства в других сценариях?

1. **Устойчивость обучения. Повышение устойчивости обучения**

Можно ввести практически незаметные для глаза человека возмущения во входные данные, которые, тем не менее, полностью поменяют выход нейронной сети. Такие почти не заметные возмущения, которые меняют выход нейросети, называют состязательными примерами (или, зачастую, атаками на эту самую нейросеть); по-английски устойчивое выражение adversarial examples / attack.

Ключевая причина состоит в недостаточной обобщающей способности нейросетей. В силу этого границы классификации, которые строятся при обучении нейросети, зачастую проходят очень близко к обучающим данным, и порой легко «заступить» с помощью минимального возмущения из области, соотвествующей одному классу, в область, соответствующую другому классу.

Частично данную проблему принято решать с помощью техники аугментации входных данных при обучении (добавление к исходным данным их немного преобразованных вариантов: например, с помощью поворота, затемнения/засветления, масштаба и пр.).

**Список источников**

1. ISO/IEC CD 24029-2 [Электронный ресурс]. URL:<https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/07/98/79804.html> (дата обращения: 12.04.2022a).

2. ISO/IEC TR 24029-1:2021 [Электронный ресурс]. URL:<https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/07/76/77609.html> (дата обращения: 12.04.2022b).

3. Robustness Metrics: How Are They Calculated, When Should They Be Used and Why Do They Give Different Results? - McPhail - 2018 - Earth’s Future - Wiley Online Library [Электронный ресурс]. URL:<https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/2017EF000649> (дата обращения: 12.04.2022).

4. Исследование устойчивости сверточных нейросетей. Часть 1: Теория [Электронный ресурс]. URL:<https://habr.com/ru/company/huawei/blog/509816/> (дата обращения: 13.04.2022).