УДК 004.852

Методы и средства обеспечения робастности нейронных сетей на основе анализа межслойной трансформации признакового пространства

Сукачев П.П. (Университет ИТМО), **Хлестунова С.Н. (Университет ИТМО)**
Научный руководитель – доцент, кандидат технических наук, Гусарова Н.Ф. (Университет ИТМО)

Введение. Нейронные сети в настоящее время находятся в фазе активного развития и интеграции в разные сферы. Этот раздел машинного обучения открывает широкие перспективы и позволяет автоматизировать совершенно разные процессы. Для применения нейросетевого подхода принципиально наличие хорошей обучающей выборки, и наличие в дефектов данных или малое их количество может существенно ухудшить результаты.

В настоящее время задача визуализации моделей машинного обучения и нейронных сетей в частности является одной из ключевых в контексте интерпретации этих моделей и исследования непосредственно процесса их обучения [1]. Для разработки качественной модели, устойчивой к шумам, можно работать не только с входными и выходными данными, но и промежуточными состояниями признакового пространства датасета. С помощью топологического подхода можно рассмотреть признаковое пространство и его эволюцию в процессе обучения [2].

.

Основная часть. Основная цель – разработка основанных на топологическом анализе данных методов для повышения устойчивости модели к различным дефектам данных. Для достижения цели выделяются следующие задачи:

1) Разработка модели с динамической визуализацией признакового пространства (вывод признакового пространства каждые n эпох).

2) Сравнение результатов в применении к «чистому» датасету и зашумленному, выбор метрики сравнения.

3) Разработка модели для обучения нейронной сети на малой выборке с применением топологического анализа.

Для решения задач сперва используются простые датасеты, чтобы изучить модели и выделить некоторые свойства. На генерированном с помощью простого алгоритма датасете, в частности, можно визуально наблюдать особенности топологических преобразований слоев с разными функциями активациями. На более сложных данных с высокоразмерным признаковым пространством (например, двумерные изображения) применяются сверточные, полносвязные и рекуррентные нейронные сети; применяются также методы кластеризации и понижения размерности, результаты сравниваются с результатами заведомо плохо обучающейся модели.

Далее можно выделить подзадачу поиска в признаковом пространстве аттракторов и их эволюцию в процессе обучения. Аттракторы можно рассматривать, как точки притяжения элементов конкретного класса, в таком случае обучение может быть представлено движением точек в признаковом пространстве в полях этих аттракторов.

Выводы. Построена модель с визуализацией топологических преобразований признакового пространства на полносвязных и рекуррентных сетя, с помощью топологического анализа данных были изучены различные архитектуры. В дальнейшем планируется разработка основанного на топологическом анализе данных метода XAI для повышения эффективности классификаторов.

.

Список использованных источников:

1. German Magai, Anton Ayzenberg, Topology and geometry of data manifold in deep learning // arXiv.org. 2022. Дата обновления: 19.02.2022. URL: https://arxiv.org/abs/2204.08624 (дата обращения: 22.11.2022).
2. Mustafa Hajij, Kyle Istvan, A Topological Framework for Deep Learning // arXiv.org. 2020. Дата обновления: 21.06..2021. URL: https://arxiv.org/abs/2008.13697 (дата обращения: 24.11.2022).