УДК 004.8

МЕТОДЫ ИССЛЕДОВАНИЯ ХАРАКТЕРИСТИЧЕСКИХ СВОЙСТВ НЕЙРОННЫХ СЕТЕЙ С ПРИМЕНЕНИЕМ ТЕОРЕТИКО-ИГРОВОГО ПОДХОДА

Зайцева М.А. (ИТМО), **Томилов И.В.** (ИТМО)
Научный руководитель – кандидат технических наук, доцент Гусарова Н.Ф.
(ИТМО)

Введение. Нейронные сети играют ключевую роль в современных системах искусственного интеллекта, демонстрируя высокую эффективность в различных областях. Однако, глубинное понимание их функционирования и теоретическое обоснование методов оптимизации остаются важными задачами. В рамках данной работы предложен метод регуляризации нейронных сетей, основанный на методе DropGrad [1] и отличающийся тем, что обнуление градиента для отдельных нейронов происходит не на этапе мета-оптимизации в задаче мета-обучения, а на этапе оценки параметров в задаче стандартного обучения с учителем. Для анализа и интерпретации полученных результатов разработана теоретическая модель на основе композиционной теории игр [2], позволяющая рассматривать слои нейронной сети как отдельных "игроков", взаимодействующих в процессе обучения. Такой теоретический подход призван углубить понимание внутренних механизмов работы нейронных сетей и способствовать созданию более интерпретируемых и устойчивых моделей.

Основная часть. В рамках данного исследования предложен новый подход к анализу устойчивости нейронных сетей, основанный на применении композиционной теории игр.

1) Нейронная сеть моделируется как композиционная игра, где каждый слой или компонент сети рассматривается как отдельный игрок. Взаимодействие между слоями формализовано как игра между агентами, стратегии которых определяются параметрами обучения и регуляризации.
2) Исследовано влияние случайного шума на входах сети на точность предсказаний в задачах классификации и регрессии. В качестве модели игровой стратегии слоя выбрана регуляризация на основе метода DropGrad [1]. Модели обучались на наборах из коллекции Penn Machine Learning Benchmarks [3] и из других открытых источинков. Экспериментальные результаты демонстрируют, что изменение параметров регуляризации, таких как вероятность выпадения градиента, влияет на устойчивость сети к шуму.

Выводы. Предложенный подход на основе композиционной теории игр предоставляет перспективные инструменты для анализа поведения нейронных сетей. Показано, что изменение параметров регуляризации может быть интерпретировано как изменение "игровой стратегии" отдельных слоёв, что позволяет оптимизировать общую устойчивость сети к шуму. Работа открывает горизонты для более глубокого понимания внутренних механизмов работы нейронных сетей.

Список использованных источников:

1. H.-Y. Tseng, Y.-W. Chen, Y.-H. Tsai, S. Liu, Y.-Y. Lin, M.-H. Yang. Regularizing meta-learning via gradient dropout // Computer Vision – ACCV 2020. 15th Asian Conference on Computer Vision, Kyoto, Japan, November 30. – 2020. – pp. 218-334.
2. N. Ghani, J. Hedges, V. Winschel, P. Zahn. Compositional game theory // Proceedings of the 33rd Annual ACM/IEEE Symposium on Logic in Computer Science, LICS. – 2018. – pp. 472-481.
3. J. D. Romano, T. T. Le, W. La Cava, J. T. Gregg, D. J. Goldberg, P. Chakraborty, N. L. Ray, D. Himmelstein, W. Fu, J. H. Moore. Pmlb v1.0: an open source dataset collection for benchmarking machine learning methods // Bioinformatics, 38(3). – 2022. – pp. 878-880.