Проверка надёжности нейронных сетей методом фаззинга

Запорожец Илья Владимирович[1]
Мишечкин Максим Владимирович[2]

1: *Студент, факультет ИУ МГТУ имени Н. Э. Баумана, Москва, Россия*
2: *Аспирант, ИСП РАН имени В. П. Иванникова, Москва, Россия*
E-mail: ilyazap@ispras.ru, mishmax@ispras.ru

Научный руководитель — Курмангалеев Шамиль Фаимович

В современном мире все больше находят свое применения различные системы распознавания на основе нейронных сетей (глубокого обучения). Они могут быть использованы для широкого спектра задач — от обнаружения преступников до работы в системах автоматического управления транспортными средствами — и должны быть в высокой степени надежными. По этой причине появилась необходимость в качественном тестировании подобных систем и своевременном выявлении уязвимостей. Одним из подходов является применение метода фаззинга к задаче тестирования нейронных сетей.

Идея фаззнга заключается в автоматической передаче случайных или изменяемых по определенному алгоритму входных данных тестируемой программе. Интересными являются те, которые приводят к падениям, зависаниям или нарушению логики работы. В некоторых фаззерах, например, из семейства AFL (American Fuzzy Loop)[4], вводится понятие покрытия для оптимизации процесса. При таком подходе фаззер «запоминает» те входные данные, которые привели к расширению покрытия, и использует их для дальнейших мутаций.

В применении к нейронным сетям используются особые метрики покрытия из-за специфики задачи. Например, предлагается получать информацию об активированных нейронах и использовать ее в качестве покрытия[1]. При этом подход авторов предполагает использование сразу нескольких нейронных сетей и поиск расхождений в их предсказаниях, что, как нам кажется, является ограничивающим фактором при тестировании. Также предметом анализа может выступать выходной слой сети, представленный в качестве вектора[2]. При таком подходе фаззер определяет, насколько далеко в пространстве данный вектор находится от уже полученных ранее, и на основе этого выносит решение о получении нового покрытия.

В рамках данной работы была реализована возможность фаззинга нейронных сетей в современном инструменте Crusher, разработанного в ИСП РАН и вобравшего в себя передовые технологии в области фаззинга[3]. В качестве метода сбора покрытия мы предлагаем использовать комбинированный метод из подходов, упомянутых выше, что позволяет компенсировать недостатки каждого из них по отдельности. При этом наш инструмент не требует наличия нескольких нейронных сетей у пользователя — для проведения тестирования необходима только одна. На данный момент реализована поддержка крупнейших фреймворков машинного обучения — TensorFlow и PyTorch — с гибкой настройкой применительно к конкретной задаче фаззинга, а благодаря архитектуре Crusher процесс легко масштабируется.

Наш подход демонстрирует высокую эффективность в фаззинге различных современных нейронных сетей. Например, на рабочей машине с десятью физическими ядрами первые результаты поиска неверной классификации сетью YOLOv8 удалось получить уже через 30 секунд после начала фаззинга. Аналогичные результаты были получены и для сети FaceNet, специализированной для задачи распознавания лиц.

Литература

1. Kexin Pei, Yinzhi Cao, Junfeng Yang, Suman Jana DeepXplore: Automated Whitebox Testing of Deep Learning Systems // arXiv:1705.06640v4. 2017
2. Odena A., Goodfellow I. TensorFuzz: Debugging Neural Networks with Coverage-Guided Fuzzing // arXiv:1807.10875. 2018
3. ИСП Crusher: комплекс динамического анализа программ: `https://www.ispras.ru/technologies/crusher/`
4. AFL technical details: `https://github.com/google/AFL/blob/master/docs/technical_details.txt`