УДК 004.8

ОПРЕДЕЛЕНИЕ АРХИТЕКТУРЫ НЕЙРОННЫХ СЕТЕЙ С ПОМОЩЬЮ ФАЗЗИНГА БИНАРИЗОВАННЫХ МОДЕЛЕЙ

Тимощук-Бондарь А.И.[1]
Научный руководитель – канд. техн. наук Кугаевских А.В. [1]
[1]Университет ИТМО
aitimoshchuk-bondar@itmo.ru

Работа выполнена в рамках темы НИР № 15012 «Определение архитектуры нейронных сетей с помощью фаззинга бинаризованных моделей».

Введение

В современных реалиях нейронные сети чаще всего используются в режиме «чёрного ящика»: исследователю или инженеру доступны только входные данные и выходные ответы модели, тогда как внутренняя архитектура скрыта. Это характерно для моделей, предоставляемых через API, для проприетарных решений, а также для ситуаций, когда требуется аудит сторонней модели без доступа к её весам и коду. Отсутствие информации об архитектуре затрудняет интерпретацию поведения сети, выбор корректных процедур тестирования, а также оценку устойчивости и корректности в нетипичных режимах работы. Существующие подходы к тестированию нейросетей в значительной степени опираются либо на белый ящик, либо на трудоёмкую ручную разметку и не всегда дают инструмент для восстановления архитектурных признаков по наблюдаемому поведению. В качестве методологической опоры в работе использованы идеи coverage-guided тестирования и фаззинга для нейросетей, предложенные в DeepXplore, TensorFuzz и DeepHunter [1-3], где вводятся понятия нейронного покрытия и направленного отбора тестов, увеличивающих разнообразие внутренних состояний модели.

Основная часть

Цель исследования - экспериментально проверить возможность восстановления архитектурных признаков однослойной нейронной сети по парам «входные данные -> выходные логиты» и предложить практический инструмент анализа. Для формирования разнообразного пула тестируемых «чёрных ящиков» реализован генератор однослойных моделей, который перебирает тип целевого слоя (CNN, полносвязный слой, RNN, LSTM) и тип функции активации (ReLU, LeakyReLU, Sigmoid, Softmax). Таким образом получен набор моделей, различающихся по вычислительной природе (локальная свёртка, глобальная линейная проекция, рекуррентная динамика) и по нелинейности. Далее построен фаззинг-пайплайн для генерации входных данных, способных «выявлять» архитектуру через наблюдаемые выходы. В качестве базового корпуса выбран MNIST как стандартный, стабильный и воспроизводимый источник входов фиксированного формата. Фаззинг организован в две стадии. На первой стадии выполняются целевые мутации, различающиеся по предполагаемому типу слоя: для CNN используются локальные геометрические и фотометрические искажения (повороты, небольшие сдвиги, структурированный шум, вариации толщины линий), для полносвязных моделей - глобальные преобразования яркости и равномерные шумовые возмущения, для RNN/LSTM - последовательностные мутации при интерпретации изображения как последовательности строк (реверс, добавление шумовых префиксов/суффиксов, локальные деформации). На второй стадии используется coverage-guided отбор: для каждой модели собираются статистики активаций, вычисляется метрика покрытия и сохраняются только те мутанты, которые расширяют наблюдаемое поведение, тогда как поведенческие дубликаты отбрасываются. Такой подход согласуется с идеологией

neuron coverage и coverage-guided fuzzing, сформулированной в DeepXplore и TensorFuzz [1, 2] и развитой в DeepHunter. В результате сформирован расширенный фазз-датасет порядка 81 400 примеров, обеспечивающий высокое разнообразие входов при контролируемом размере выборки. Задача восстановления архитектуры формализована как multi-label классификация по логитам. Архитектурные признаки кодируются 8-битным двоичным вектором: первые 4 бита соответствуют типу слоя (CNN, FC, RNN, LSTM), последние 4 - типу активации (ReLU, LeakyReLU, Sigmoid, Softmax). В ходе исследования мной были предложены два метода анализа: (1) трансформер-декодер, обрабатывающий логиты как последовательность и предсказывающий 8 выходных узлов; (2) MLP, работающий по агрегированным статистикам логитов. В основе первого метода лежат принципы self-attention, описанные в работе Attention Is All You Need [4]. В экспериментальной оценке трансформер-декодер показал устойчивое качество на задаче восстановления архитектурных битов (mAP = 0.92; Macro F1 по типам слоёв около 0.87; Macro F1 по типам активаций около 0.84), при этом наибольшая доля ошибок связана с различением ReLU и LeakyReLU как близких по форме нелинейностей и, следовательно, по распределениям логитов. Второй метод в ходе исследования показал свою несостоятельность.

Выводы

В результате исследования мной предложен и реализован полный воспроизводимый контур анализа архитектуры однослойной нейросети в условиях «чёрного ящика»: генерация контролируемого набора моделей, построение coverage-guided фазз-датасета, фиксация выходных логитов и обучение классификаторов, предсказывающих тип слоя и тип функции активации. Практическая ценность результата заключается в возможности аудита и профилирования сторонних моделей без доступа к их внутреннему устройству: подход применим для задач тестирования надёжности, выбора корректных фазз-мутаций под предполагаемый тип сети, а также для задач верификации соответствия заявленных характеристик модели реальному поведению. В результате повышается доверие к результатам работы моделей, а также уверенность в отсутствии непредсказуемых результатов при получении нестандартных входных данных в ходе эксплуатации или состязательных атак.

Литература

1. Pei K., Cao Y., Yang J., Jana S. DeepXplore: Automated Whitebox Testing of Deep Learning Systems // Proceedings of the 26th Symposium on Operating Systems Principles. - 2017. - P. 1–18. - DOI: 10.1145/3132747.3132785.
2. Odena A., Goodfellow I. TensorFuzz: Debugging Neural Networks with Coverage-Guided Fuzzing. - 2018. - DOI: 10.48550/arXiv.1807.10875.
3. Xie X., Ma L., Juefei-Xu F., Chen H., Xue M., Li B., Liu Y., Zhao J., Yin J., See S. DeepHunter: Hunting Deep Neural Network Defects via Coverage-Guided Fuzzing. - 2018. - DOI: 10.48550/arXiv.1809.01266.
4. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I. Attention Is All You Need. - 2017. - DOI: 10.48550/arXiv.1706.03762.