УДК 004.415.5, 004.896

РАЗРАБОТКА МЕТОДОВ ФОРМАЛЬНОЙ ВЕРИФИКАЦИИ НЕЙРОСЕТЕВЫХ КОНТРОЛЛЕРОВ В ЗАДАЧАХ МАРШРУТИЗАЦИИ

Давлетшин Р.О., Университет ИТМО

Научный руководитель – Чивилихин Д.С., кандидат технических наук, научный сотрудник факультета ИТиП Университета ИТМО

В данной работе предлагаются методы формальной верификации нейросетевых контроллеров в задачах маршрутизации, основанные на сведении к задаче о выполнимости булевой формулы.

Введение. Большинство существующих методов решения задач маршрутизации исходят из предположения, что граф, для которого решается задача маршрутизации, фиксирован. Однако на практике это условие зачастую не выполняется, поскольку реальные системы подвержены регулярным воздействиям, меняющим топологию или свойства графа. В подобных случаях либо выбранный алгоритм применяется заново для каждого нового графа, либо продолжает использоваться решение, построенное для исходного графа. Для работы с такими условиями в последнее время всё более популярными становятся подходы к решению задачи маршрутизации в мультиагентной парадигме, где перемещаемый объект является агентом, для управления которым используются алгоритмы на основе глубокого обучения с подкреплением. Как и во множестве других приложений глубокого обучения с подкреплением, решения, принимаемые с помощью нейронных сетей, могут влиять на безопасность и здоровье людей. Поэтому особую актуальность приобретает вопрос корректности работы нейронных сетей, однако в отличие от традиционных программ, написанных на языках программирования, нейронную сеть принципиально невозможно проверить вручную. Вследствие чего актуальна разработка методов автоматической верификации заданных свойств нейронных сетей, методов ее обучения и синтезируемых ей решений задачи маршрутизации.

Основная часть. Методы верификации свойств нейронных сетей получили наибольшее развитие в области классификации изображений. Интерес к данному приложению обусловлен развитием и реализацией технологий беспилотных автомобилей, которые во многом зависят от методов распознавания изображений, получаемых от камер, установленных на автомобиле. В основном в работах исследователей рассматривается задача проверки устойчивости классификации в присутствии возмущений - заданных типов манипуляций с изображением. Целью таких методов верификации является проверка того, что результат классификации изображений, визуально незначительно отличающихся от заданного, совпадает с результатом классификации этого заданного изображения. Стоит отметить, что наибольшую ценность представляют надежные, полные результаты верификации. Таким образом если контрпример существует, то он должен быть найден. Если же верификация сообщает, что контрпримера нет и нейронная сеть надежна — это должно соответствовать истине. Поэтому большинство методов верификации основываются на сведении к задачам выполнимости булевой формулы. Одним из вариантов такого сведения является кодирование структуры и активации сети с помощью логических переменных и аппроксимации кусочно-постоянными функциями нелинейных зависимостей. Отдельно выделяют другое направление исследований — верификацию бинаризованных глубоких нейронный сетей. За счет того, что веса и активации таких сетей являются бинарными, задачи верификации для таких сетей эффективно сводятся к задаче о выполнимости булевой формулы.

Методы, разработанные в рамках данной работы, основываются на подходах по верификации нейронных сетей в области классификации изображений и решают задачу формальной проверки целевого нейросетевого контроллера на предмет обеспечения доставки пакета между заданными вершинами графа при постоянной топологии.

Выводы. В результате работы были разработаны новые методы формальной верификации нейросетевых контроллеров в задачах маршрутизации, основанные на сведении к задаче о выполнимости булевой формулы, для случая фиксированной топологии графа маршрутов и фиксированных точек начала и конца движения пакета. Также было проведено экспериментальное исследование предложенных методов, которое показало эффективность предложенных методов.

Давлетшин Р.О. (автор) Подпись

Чивилихин Д.С. (научный руководитель) Подпись