УДК 004.932.2+004.932.72'1

ОБЛАЧНЫЙ СЕРВИС ПОЛУАВТОМАТИЧЕСКОГО ФОРМИРОВАНИЯ НАБОРОВ ДАННЫХ ДЛЯ ОБУЧЕНИЯ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ

Фомин И.С. (ГНЦ РФ Центральный научно-исследовательский и опытно-конструкторский институт робототехники и технической кибернетики (ЦНИИ РТК)), **Филатов Н.С.** (ГНЦ РФ Центральный научно-исследовательский и опытно-конструкторский институт робототехники и технической кибернетики (ЦНИИ РТК))

Научный руководитель – к.т.н, доцент Бахшиев А.В.

(ГНЦ РФ Центральный научно-исследовательский и опытно-конструкторский институт робототехники и технической кибернетики (ЦНИИ РТК), Санкт-Петербургский политехнический университет Петра Великого (СПбПУ))

На сегодняшний день нейронные сети показывают высокое качество в задачах анализа сенсорной информации, в особенности для систем технического зрения. Для обучения сетей требуются наборы данных, содержащие большое количество изображений. Наборы данных большого объема невозможно создавать только с помощью ручной разметки, временные и человеческие затраты слишком велики. Поэтому разработка технических решений для автоматизации формирования наборов данных и их валидации является важной перспективной задачей.

Введение. На сегодняшний день многие задачи технического зрения, включая классификацию изображений, обнаружения и семантической сегментации объектов находят свои решения с использованием сверточных нейронных сетей. Для большинства условий качество решения задач глубокими нейронными сетями выше, чем для не-нейросетевых методов. Все алгоритмы машинного обучения требуют для решения конкретной задачи наборы данных с соответствующей разметкой. Для того, чтобы процедура обучения алгоритма прошла успешно, набор должен обладать целым рядом особых качеств. Набор должен быть репрезентативным, то есть содержать примеры для всех условий с которыми столкнется алгоритм по окончании обучения. Количество примеров должно быть достаточно большим чтобы нейронная сеть обучилась распознавать все необходимые случаи без переобучения (т.е. простого «запоминания» поданных примеров). В наборе должны соблюдаться баланс сложности примеров. Исторически, создание размеченных (аннотированных) наборов данных началось одновременно с появлением алгоритмов машинного обучения, поначалу небольших. С появлением сверточных сетей выросло качество распознавания (классификации, обнаружения, сегментации) объектов на изображениях. Но также существенно выросло количество настраиваемых параметров сети (весов) и, следовательно, количество требуемых для обучения данных. Для первых сверточных архитектур это количество составляло десятки и сотни тысяч примеров. Разметка такого объема данных осуществляется частично самими исследователями, частично – с помощью краудфандинговых систем (Amazon Mechanical Turk и его аналоги), в которых пользователи получают небольшие суммы за выполнение заданий по аннотированию. С ростом размера сетей, сложности и количества слоев, выросло количество требуемых данных. Современные наборы для обнаружения и классификации содержат десятки миллионов изображений. Размечать такой объем вручную для новой задачи - слишком долго и дорого, поэтому активно применяются уже обученные сети для обнаружения и классификации с последующей верификацией, к примеру, через AMT. Для наборов данных для сегментации пока не создано методов качества столь высокого чтобы можно было применять их для автоматизации аннотирования, большинство размечаются вручную.

Основная часть. Сложности с формированием собственных наборов данных рано или поздно возникают у всех групп исследователей, которые работают с нейронными сетями. В

зависимости от задачи формирование разметки может требовать большого количества усилий. Вместе с этим среди методов машинного обучения существуют средства, способные обеспечить помощь при формировании аннотаций для классификации и обнаружения объектов на изображениях. В случае сегментации ситуация более сложная, но и в этом случае методы машинного обучения способны оказать существенную помощь, например, в виде выделения регионов без определения классов.
Средство для помощи в аннотации данных должно быть облачным, только таким способом можно организовать использование данных, получаемых разными группами разработчиков для улучшения качества работы алгоритмов. Распределенная структура, включает в себя глобальный сервер для связи экземпляров клиентских и серверных приложений, запущенных на разных ПК, распределенную базу данных, объединяющую информацию о наборах данных, пользователях, существующих архитектурах и наборах весов для них, а также хранилище данных, включающее готовые к разворачиванию на расчетной машине наборы данных, архитектуры, веса. Клиентская часть системы включает приложение, позволяющее просматривать имеющиеся архитектуры с готовыми наборами весов и определять их возможности по формированию аннотаций. Также клиентское приложение позволит выполнять удаленную настройку и разметку данных, как в полностью ручном режиме, так и с поддержкой методов, упрощающих ручную разметку; обеспечивать ручное исправление автоматически сформированных аннотаций, или валидацию для исправления ошибок классификации и окончательной подготовки данных к обучению сети. Алгоритмы оценки позволят оценить качество и репрезентативность набора при условии наличия данных, на которых будет выполняться тестирование. Хранящаяся в системе информация о том, как параметры алгоритма и набора данных связана с качеством сети на конкретном наборе данных после обучения в перспективе позволит создать алгоритмы формирования рекомендаций по применению той или иной архитектуры в конкретной задаче.

Выводы. Возможность сформировать набор данных, содержащий только те, которые необходимы для решения конкретной практической задачи, обучить любую архитектуру из представленных в системе на таком наборе данных используя удобный пользовательский интерфейс или получить готовое решение в виде подготовленной к решению задачи и оптимизированной для сокращения затрат нейронной сети с набором весов принесет существенную пользу многим исследователям. Также использование такой системы понизит порог вхождения для людей, ранее возможно не задумывавшихся о применении методов глубокого обучения для своих задач ввиду практической сложности освоения этой технологии и потребности в дорогостоящем оборудовании для обучения. Возможность размечать данные в ручном, полуавтоматическом и автоматическом режимах увеличит поток новых данных от различных команд исследователей, что позволит существенно расширить количество данных и весов в системе, повышая ее полезность.

Фомин И.С. (автор) Подпись

Филатов Н.С. (соавтор) Подпись

Бахшиев А.В. (научный руководитель) Подпись