



Contribution ID: 144

Type: Sectional reports

Обучение без учителя на данных узкой направленности агрегированных автоматически.

Tuesday 6 July 2021 15:50 (15 minutes)

Информация, публикуемая пользователями в открытом доступе, может служить хорошим ресурсом для сбора данных при формировании датасетов для обучения нейронных сетей. Одной из самых больших существующих платформ для обмена фотографиями и видеозаписями является Instagram. Основным методом взаимодействия пользователей друг с другом на данной платформе является публикация изображений. При этом доступны такие функции как: добавление описания к изображению, хэштег, отметка рамкой на изображении и т.д. Это обстоятельство позволяет исследователям использовать Instagram для применения методов машинного обучения и анализа изображений.

Данная платформа обладает рядом особенностей, которые могут быть использованы при создании обучающих датасетов для нейросетевых моделей, работающих с изображениями. С помощью хэштегов можно получать наборы размеченных данных по заданной тематике. Кроме того, метаданные, получаемые вместе с изображением, могут быть полезны для поиска связанных с тематикой профилей, уточнения разметки и автоматического взаимодействия с пользователями.

Получение данных с платформы Instagram возможно несколькими общеизвестными способами: последовательность запросов к официальному API платформы, использование инструментов автоматизации работы веб-браузера и всевозможных веб-краулеров. Эксплуатация официального API обладает рядом ограничений, описанных в документации. Меняющийся интерфейс Instagram делает затруднительным применение таких инструментов автоматизации работ веб-браузера, как Selenium. В то время как использование веб-краулера показывает свою эффективность для сбора общедоступных изображений, а также связанных с ними метаданных.

Полученные таким образом базы данных, после предварительной обработки могут быть использованы в качестве обучающих датасетов для нейронных сетей. Предварительная обработка такого рода датасетов необходима, поскольку данные могут быть сильно разрознены. С помощью моделей классификации возможно очистить датасет от ошибок "разметки", допущенных пользователями в описаниях к публикуемым изображениям.

Нашей целью является создание автоматического приложения, способного создавать тематические датасеты по хэштегам и профилям. В функционал приложения также входит отслеживание хэштега и автоматическая публикация ответного сгенерированного изображения. Генерация осуществляется с помощью наложения маски, полученной нейронной сетью для сегментации объектов, однако, для улучшения качества сегментации, предварительно используется super-resolution модель.

Проблема неточности разметки, полученной из метаданных, была решена с помощью self-supervised обучения модели. Применение метода DINO (self-distillation with no labels) с нейронной сетью, в основе которой лежит трансформер, способствовало улучшению качества сегментации.

В результате была создана система, способная реагировать на события платформы Instagram. Ответ пользователю формируется с помощью нейронных моделей и автоматически публикуется.

Summary

Authors: Mr IAKUSHKIN, Oleg (Saint Petersburg State University); PAVLOVA, Ekaterina (Saint Petersburg State University)

Presenter: PAVLOVA, Ekaterina (Saint Petersburg State University)

Session Classification: Round table on IT technologies in education

Track Classification: 9. Big data Analytics and Machine learning