Библиотека Конгресса США обрабатывает более 16 миллионов исторических газетных страниц с использованием ИИ

Оцифровка миллионов исторических документов и газет является сложной задачей. Чтобы ускорить процесс, Библиотека Конгресса США разработала модель глубокого обучения с ускорением на графическом процессоре для автоматического выделения, категоризации и подписи более 16 миллионов страниц исторических американских газет, опубликованных в период с 1789 по 1963 годы.

Работа, которая становится общедоступной, включает в себя визуальный контент, заголовки, фотографии, иллюстрации, карты, комиксы, редакционные мультфильмы и рекламные объявления из исторических газет. По словам исследователей, набор данных является самым большим из когда-либо созданных. Эта работа является частью инициативы организации «Хроника Америки», которая вытекает из партнерства между Библиотекой Конгресса и Национальным гуманитарным фондом.

«Более 16 миллионов страниц исторических американских газет были оцифрованы для Chronicling America на сегодняшний день, в комплекте с изображениями с высоким разрешением и машиночитаемыми METS / ALTO и оптическим распознаванием символов (OCR)», говорится в статье. Набор данных: извлечение и анализ визуального контента из 16 миллионов исторических газетных страниц в хронике Америки. METS и ALTO – это стандарты XML, поддерживаемые Библиотекой Конгресса, которые включают локализацию текста.

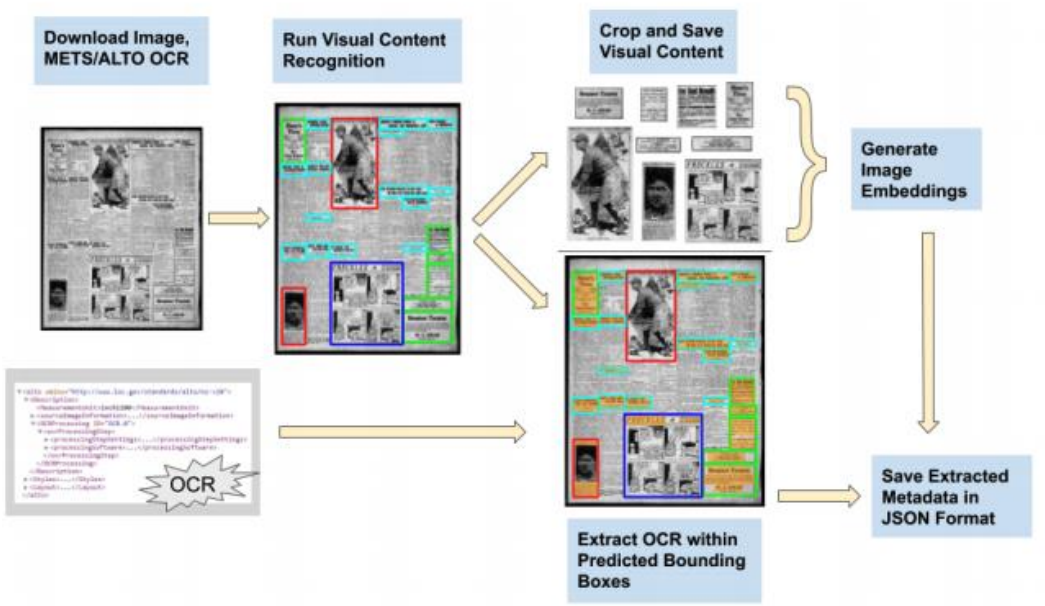

Рисунок. Этапы распознавания.

Помимо выпуска нового набора данных, исследователи также опубликовали свою модель распознавания визуального контента для исторических газет, а также новый учебный набор данных для этой задачи. Модели основаны на краудсорсинговых инициативах Библиотеки по аннотированию и подписанию визуального контента в газетах времен 1-й мировой войны.

Обучение и выводы проводились на графических процессорах NVIDIA с ускоренной структурой глубокого обучения PyTorch с ускорением cuDNN. Все тонкие настройки выполнялись с использованием графических процессоров NVIDIA T4 в облаке Amazon Web Services с использованием магистрали R50-FPN. Исследователи объяснили, что эта магистраль была выбрана потому, что она имеет самое быстрое время вывода из магистралей с более быстрым RCNN.

Источник: https://news.developer.nvidia.com/u-s-library-of-congress-processes-over-16-million-historic-newspaper-pages-using-ai/