УДК 004.932

Внедрение современных инструментов инференса для нейронных сетей в сервис исправления смазанных фотографий Deblur

Султанов М.М. (ИТМО)
Научный руководитель – кандидат технических наук, доцент Лукин М.А.
(ИТМО)

Введение. Современные алгоритмы обработки изображений широко используют нейросетевые модели для восстановления качества фотографий, в том числе для устранения смазанности, возникающей вследствие движения камеры или объекта [1–3]. В данной работе основное внимание уделяется улучшению кодовой базы существующего проекта по исправлению смазанности изображений. В настоящее время для инференса отдельных нейросетевых моделей, реализованных преимущественно на PyTorch, используется разрозненный набор решений. Цель работы — внедрить универсальную платформу для инференса нейросетей, которая позволит объединить и оптимизировать имеющийся функционал, обеспечив кроссплатформенную совместимость (Windows, Linux) и возможность использования как GPU (Nvidia), так и CPU (при отсутствии видеокарты с соответствующими уведомлениями для пользователя). Приведение моделей к стандартизированному формату ONNX [4,5] и дальнейшая оптимизация инференса с использованием TensorRT [6] являются ключевыми инструментами для достижения поставленной цели.

Основная часть. Для реализации цели работы решаются следующие задачи:

1. Унификация инференс-кода.
   Объединение разрозненных реализаций инференса нейросетевых моделей, выполненных на PyTorch, в единую кодовую базу.
2. Конвертация моделей в формат ONNX.
   Приведение существующих моделей к формату ONNX для обеспечения стандартизации и кроссплатформенной совместимости [4,5].
3. Оптимизация инференса.
   Применение TensorRT для ускорения работы модели и снижения вычислительных затрат, что особенно актуально при работе с большими объёмами данных [6].
4. Кроссплатформенная поддержка.
   Обеспечение работы системы как на Windows, так и на Linux с возможностью выбора между GPU и CPU в зависимости от доступного оборудования.
5. Разработка механизма управления задачами.
   Создание системы управления очередью задач для параллельного выполнения запросов, что позволяет эффективно использовать вычислительные ресурсы при высокой нагрузке.

Выводы. В результате проведённой работы будет создана унифицированная платформа для инференса нейросетевых моделей, позволяющая значительно ускорить обработку изображений за счёт оптимизации вычислительных процессов и упрощения развертывания решения. Использование стандартизированного формата ONNX [4,5] и оптимизация с помощью TensorRT [6] обеспечат высокую производительность и масштабируемость системы, делая проект пригодным для применения как в персональном, так и в многопользовательском режиме.

Список использованных источников:

1. Fergus R., Singh B., Hertzmann A., Roweis S., Freeman W. Removing camera shake from a single photograph // ACM Transactions on Graphics. — 2006. — Т. 25, №3. — С. 787–794. — DOI: 10.1145/1179352.1141956.
2. Chen L., Zhang J., Lin S., Fang F., Ren J. S. Blind Deblurring for Saturated Images // 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). — 2021. — С. 6304–

6312. — DOI: 10.1109/CVPR46437.2021.00624.

3. Kupyn O., Budzan V., Mykhailych M., Mishkin D., Matas J. DeblurGAN: Blind Motion Deblurring Using Conditional Adversarial Networks // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2018. — С. 8183–8192. — DOI: 10.1109/CVPR.2018.00854.

4. Microsoft. Преобразование модели обучения PyTorch в формат ONNX [Электронный ресурс]. — URL: https://learn.microsoft.com/ru-ru/windows/ai/windows-ml/tutorials/pytorch-convert-model (дата обращения: 25.02.2025).

5. ONNX. Get Started [Электронный ресурс]. — URL: https://onnx.ai/get-started.html (дата обращения: 25.02.2025).

6. NVIDIA. TensorRT Developer Guide [Электронный ресурс]. — URL: https://docs.nvidia.com/deeplearning/tensorrt/archives/tensorrt-803/developer-guide/index.html (дата обращения: 25.02.2025).