Национальный исследовательский
Нижегородский государственный университет им. Н.И. Лобачевского
Институт информационных технологий, математики и механики

Учебный курс «Тензорные компиляторы для глубоких нейросетевых моделей»

Ю.А. Родимков, Е.П. Васильев,
И.Б. Мееров, А.В. Сысоев, В.Д. Кустикова

Терминология

- ***Вывод (inference)*** – прямой проход по обученной нейронной сети с целью решения задачи

- ***Тензоры (tensors)*** – многомерные матрицы данных и параметров модели, над которыми выполняются операции в процессе вывода

- ***Тензорные компиляторы (tensor compiler)*** – инструменты, которые предназначены для ускорения вывода нейронных сетей за счет оптимизации модели на разных уровнях представления и компиляции под конкретное устройство

Цели

- ***Цель курса*** – понять общее устройство тензорных компиляторов и научиться использовать инструментарий для оптимизации вывода нейросетевых моделей при решении практических задач

Принципы построения курса

- Самодостаточность разрабатываемых материалов
- Ориентированность на практическое применение тензорных компиляторов

Целевая аудитория и требования

- ***Целевая аудитория:***
  - Студенты старших курсов и аспиранты технических и естественно-научных специальностей
  - Инженеры ИТ-компаний
  - Преподаватели вузов и исследователи

- ***Требования к слушателям:***
  - Базовые навыки разработки программ на языках C, C++, Python
  - Знания и навыки в дискретной математике, линейной алгебре, системном программировании, алгоритмах и структурах данных

Структура курса

- ***Структура курса:***
  - 16 часов лекций (7 лекций продолжительностью от 1 до 5 академических часов в зависимости от сложности темы)
  - 20 часов практических занятий (4 задания)

- ***Лекция*** – классическая лекция с элементами мастер-класса

- ***Практика*** – запуск и пошаговая оптимизация открытых моделей с целью достижения наилучших показателей производительности вывода

Лекционная часть курса

1	***Лекция 1.*** Введение в глубокое обучение ***Лекция 2.*** Основные виды нейросетевых моделей ***Лекция 3 (в формате мастер-класса).*** Обзор инструментов глубокого обучения, поддерживающих обучение и тестирование глубоких моделей (PyTorch, OpenVINO toolkit, TensorFlow lite)
2	***Лекция 4.*** Общая схема построения тензорных компиляторов ***Лекция 5.*** Обзор существующих тензорных компиляторов
3	***Лекция 6.*** Обзор Apache TVM ***Лекция 7.*** Автоматическая оптимизация вывода нейросетей на уровне слоев в Apache TVM

Практическая часть курса

x86

Практика 1. Программная реализация вывода глубоких нейронных сетей для классификации изображений средствами нескольких фреймворков (OpenVINO toolkit, TensorFlow lite)

x86, RISC-V

Практика 2. Вывод глубокой модели средствами Apache TVM

RISC-V

Практика 3. «Ручная» оптимизация оператора в рамках Apache TVM на примере полносвязного и сверточного слоя

Практика 4. Автоматическая оптимизация моделей в рамках Apache TVM

Заключение

- Разрабатываемый курс является актуальным и практически значимым
  - Глубокие нейронные сети используются повсеместно
  - Внедрение глубоких моделей требует систематизации знаний и процессов анализа качества и производительности вывода
- Лекционные материалы, практические работы 1 и 2 полностью подготовлены (практики 3 и 4 в разработке)
- Материалы курса к концу календарного года планируется опубликовать на сайте ННГУ под лицензией Apache 2.0
- Перспективы использования:
  - Спецкурс
  - Повышение квалификации
  - Молодежные школы для студентов и аспирантов