Contribution ID: **36** Type: **not specified**

SimplexLoRA

Sunday, 18 May 2025 16:06 (12 minutes)

В современном мире активно используются языковые модели и не менее важно их правильное дообучение (fine-tuning), например, техника low rank adaptation (LoRA), которая добавляет к выделенным слоям тренируемые параметры. Однако LoRA требует много памяти для достижения точных результатов, потому что на все слои добавляются адаптеры одинаковых рангов, и интуиции, на какие слои ее необходимо добавлять. В данной статье мы презентуем новый метод дообучения больших языковых моделей SimplexLoRA, который обходит обе эти проблемы с помощью адаптивного изменения рангов адаптеров. В результате работы метода ранги адаптеров на самых важных слоях становятся больше, а количество обучаемых параметров не изменяется. Мы провели эксперименты на бенчмарке GLUE, которые демонстрируют эффективность нашего подхода, и представим результаты на защите работы.

Primary author: SHALYGIN, Igor (MIPT)

Co-authors: VEPRIKOV, Andrey (Moscow Institute of Physics & Technology (MIPT)); DAVYDENKO, Grigorii (Moscow Institute of Physics & Technology (MIPT))

Presenter: SHALYGIN, Igor (MIPT)

Session Classification: 18-Машинное обучение и нейросети

Track Classification: Машинное обучение и нейросети