Байесовская дистилляция моделей на базе трансформеров

A Preprint

Игорь Н. Игнашин*
Кафедра интеллектуальныз систем
Национальный исследовательский университет «Московский физико-технический институт»
Россия, 141701, г. Долгопрудный, Институтский пер., д. 9
ignashin.in@phystech.edu

Elias D. Striatum
Department of Electrical Engineering
Mount-Sheikh University
Santa Narimana, Levand
stariate@ee.mount-sheikh.edu

15.12.2023

Abstract

В данной работе исследовано несколько способов дистилляции моделей на базе трансформеров, а также модели RNN. В статье в первую очередь уделяется внимание дистилляции посредством удаления слоя Attention из моделей. В качестве базовых моделей для дистилляции взяты модель RNN с Attention и модель трансформера seq2seq для задачи перевода текстов.

Ожидается лучшая сходимость, лучшее качество у дистиллированной модели по сравнению с моделью той же структуры, но с параметрами инициализации из произвольного нормального распределения.

Keywords Байесовская дистилляция · Трансформеры · RNN

1 Introduction

2 Headings: first level

2.

2.1 Headings: second level

2.1.1 Headings: third level

Paragraph

Список литературы

*