



МОЛБА

Школска 2023/24. година

ПОДАЦИ О ПОДНОСИОЦУ МОЛБЕ

Презиме:	Поп-Јованов	Име:	Иван
Е-пошта:	ivanpopjovanov@yahoo.com		
Број индекса:	1066/2022	Статус:	Самофинансирање
Ниво студија:	Мастер академске студије		

Тип: пријава теме мастер рада

Број молбе: м1066/2022-2

Катедра: Катедра за рачунарство и информатику

Технике квантизације језичких модела

САДРЖАЈ МОЛБЕ:

Значај теме и области:

Велики језички модели представљају један од најзначајнијих фронтова развоја модерне вештачке интелигенције и машинског учења. Раст њиховог квалитета, праћен је и растом њихове величине (мерене бројем параметара који их одређују). Није неуобичајено да потребна меморија за складиштење оваквих модела достигне неколико десетина, па и стотина гигабајта, што представља проблем за њихову ефикасну употребу у пракси. Квантизација неуронских мрежа има за циљ да значајно смањи меморијску захтевност оваквих модела и тиме олакша њихово складиштење и да повећа брзину њиховог извршавања, а без значајног смањења квалитета генерисаног текста. Овакве технике су свеprisутне у практичним применама, али је њихово унапређивање и даље активна тема истраживања.

Специфични циљ рада:

Циљ овог рада је преглед литературе и имплементација најпопуларнијих техника квантизације језичких модела. Две главне групе техника квантизације су технике које се примењују након завршетка тренинга и технике које се користе у току самог тренинга. Представници обе групе ће бити заступљени међу имплементираним и евалуираним техникама. Евалуиране технике ће бити поређене по величини модела који се добијају квантизацијом, броју токена које такви модели могу да генеришу у секунди, и квалитету генерисаног текста. Имплементација ће бити урађена у програмском језику Python користећи библиотеку PyTorch.

Комисија:

1. Ментор: Младен Николић
2. Председник: Јована Ковачевић
3. Члан: Александар Картељ

ПРИЛОЗИ МОЛБИ:

1. образац молбе за одобрење теме мастер рада

Београд, 26. мај 2024.

(потпис)

(Попуњава надлежна особа)

☒ ОДОБРАВА СЕ

☐ НЕ ОДОБРАВА СЕ

29. мај 2024., Мирослав Марић

(датум, име, презиме и потпис)

ОБРАЗЛОЖЕЊЕ: