КОМИСИЈИ ЗА СТУДИЈЕ II СТЕПЕНА ЕЛЕКТРОТЕХНИЧКОГ ФАКУЛТЕТА У БЕОГРАДУ

Комисија за студије II степена, Електротехничког факултета у Београду, на својој седници одржаној 07.09.2021. године именовала нас је у Комисију за преглед и оцену мастер рада дипл. инж. Дамјана Илишковића под насловом „Методе квантизације конволуционих неуралних мрежа". Након прегледа материјала Комисија подноси следећи

И З В Е Ш Т А Ј

1. Биографски подаци кандидата

Дамјан Илишковић је рођен 22.11.1997. године у Бањалуци. Завршио је основну школу „Никола Тесла" у Прњавору. Уписао је гимназију у Прњавору, коју је завршио са одличним успехом, као носилац Вукове дипломе и као ученик генерације. Током школовања освојио је више првих и других награда на републичким такмичењима из физике и математике и учествовао на Међународној олимпијади из физике 2016. године. Електротехнички факултет уписао је 2016. године. Дипломирао је на одсеку за Сигнале и системе 2020. године са просечном оценом 9,43. Дипломски рад одбранио је у септембру 2020. године са оценом 10. Дипломске академске – мастер студије на Електротехничком факултету у Београду, на Модулу за сигнале и системе уписао је у октобру 2020. године. Положио је све испите са просечном оценом 10.

2. Извештај о студијском истраживачком раду

Кандидат Дамјан Илишковић је као припрему за израду мастер рада урадио истраживање релевантне литературе из области теме мастер рада. Анализирана су постојећа решења и проблеми у области машинског учења, конкретно квантизација конволуционих неуралних мрежа. Истраживањем области утврђено је да постоје једноставне квантизационе методе као што су асиметрична и симетрична квантизација, које се често користе у овој области. Анализом је утврђено да наведене методе уз одређене модификације представљају добар избор за квантизацију различитих модела конволуционих неуралних мрежа.

3. Опис мастер рада

Мастер рад обухвата 53 стране са укупно 25 слика, 7 табела и 25 референци. Рад садржи увод, 3 поглавља и закључак (укупно 5 поглавља) и списак коришћене литературе.

Прво поглавље представља увод у коме су описани предмет и циљ рада. Представљени су најчешће коришћене хардверске платформе на којима се извршавају модели машинског учења и назначени начини употребе и пројектовања таквих платформи, те на основу тога и значај саме квантизације неуралних мрежа.

У другом поглављу је дата комплетна методологија рада. На почетку је дат преглед основних коцепата неуралних мрежа и конволуционих неуралних мрежа. Представљени су скупови података и модели конволуционих неуралних мрежа коришћени у раду. Предтављене су имплементационе специфичности свих карактеристичних слојева за коришћене моделе помоћу модула *NumPy* програмског језика *Python.* Детаљно су описане методе асиметричне и симетричне квантизације. Уведен је појам грануларности. Описан је значај и начин калибрације података. Детаљно је описана имплементација квантизованих слојева и разлике у конфигурацији које се уводе при квантизацији слојева разних модела. На самом крају овог поглавља детаљно је описан експеримент којим се овај рад бави и чији резултати ће бити коришћени за даљу анализу.

У трећем поглављу се износе резултати експеримента. Конкретно изнесене су тачности модела квантизованих асиметрично, симетрично и под претпоставкама различитих грануларности уз услове екперимента описане у претходном поглављу. Изнесене су и меморијске уштеде при квантизацији.

Четврто поглавље представља подробну анализу добијених резултата и могућих модификација коришћених метода квантизације. Изнесене су и неке напредније модерне методе корисне за квантизацију.

У петом поглављу су резимирани закључци добијени на основу резултата екперимента трећег поглавља и додатних екперимената кроз дискусију у четвртом поглављу. На основу ових закључака добијена је оријентациона процедура коју би требало пратити да би квантизација модела конволуционих неуралних мрежа била што успешнија. Дати су и предлози за додатна побољшања процедуре које ће бити интересантна у будућности.

4. Анализа рада са кључним резултатима

Мастер рад дипл. инж. Дамјана Илишковића се бави проблематиком квантизације конволуционих неуралних мрежа, а нарочито анализом квантизационих метода и имплементацијом квантиозованих модела конволуционих неуралних мрежа. Кванитизација модела машинског учења налази велику примену у наменским рачунарским системима данашњице, конкретно, омогућава да се већи модели без великих губитка тачности прилагоде чиповима релативно скромних могућности, уз убрзања закључивања и уштеде меморијских ресурса. Методе квантизације су имплементиране у модулу *NumPy* програмског језика *Python*, као и сами квантизовани и неквантизовани модели конволуционих неуралних мрежа.

Основни доприноси рада су: 1) приказ метода квантизације и њихова имплементација; 2) имплементација свих потребних квантизованих слојева коришћених у конволуционим моделима; 3) детаљна анализа метода квантизације и модификација тих метода; 4) обједињени закључци анализе квантизационих метода у оријентациону квантизациону процедуру; 5) могућност наставка рада на побољшању самих метода, односно могућност модификације добијене процедуре модерним методама ради додатног унапређења.

5. Закључак и предлог

Кандидат Дамјан Илишковић је у свом мастер раду успешно решио проблем квантизације конволуционих неуралних мрежа и детаљно анализирао аспекте квантизационих метода. Предложена анализа, модификације и сама квантизациона процедура могу значајно да убрзају извршавање, уштеде меморијске ресурсе и очувају оргиналну тачност коноволуционих неуралних мрежа. Кандидат је исказао самосталност и систематичност у своме поступку као и иновативне елементе у решавању проблематике овог рада.

На основу изложеног, Комисија предлаже Комисији за студије II степена Електротехничког факултета у Београду да рад дипл. инж. Дамјан Илишковић прихвати као мастер рад и кандидату одобри јавну усмену одбрану.

Београд, 10.09.2021. године

Чланови комисије:

Предраг Тадић

др Предраг Тадић, доцент

др Горан Квашчев, венредни професор