

КОМИСИЈИ ЗА СТУДИЈЕ II СТЕПЕНА ЕЛЕКТРОТЕХНИЧКОГ ФАКУЛТЕТА У БЕОГРАДУ

Комисија за студије II степена, Електротехничког факултета у Београду, на својој седници одржаној 23.06.2026 године именовала нас је у Комисију за преглед и оцену мастер рада дипл. инж. Жељко Урошевић под насловом „Оптимизација закључивања великих језичких модела кроз квантизацију модела и компресију KB кеша". Након прегледа материјала Комисија подноси следећи

ИЗВЕШТАЈ

1. Биографски подаци кандидата

Жељко Урошевић је рођен 11.1.2002. године у Смедеревској Паланци. Завршио је основну школу „Дуде Јовић” у Симићеву као вуковац. Уписао је Пожаревачку гимназију коју је завршио као вуковац. Електротехнички факултет уписао је 2020. године. Дипломирао је на одсеку за Софтверско инжењерство 2024. године са просечном оценом 9,74. Дипломски рад одбранио је у септембру 2024. године са оценом 10. Дипломске академске – мастер студије на Електротехничком факултету у Београду, на модулу за Софтверско инжењерство уписао је у октобру 2024. године. Положио је све испите са просечном оценом 10.

2. Извештај о студијском истраживачком раду

Кандидат Жељко Урошевић је као припрему за израду мастер рада спровео истраживање релевантне литературе и постојећих решења у области ефикасног закључивања великих језичких модела, са фокусом на меморијски трошак кључ-вредност кеша (енгл. Key-Value Cache, KB кеш). Проучени су трансформер архитектура и варијанте механизма пажње, значај малих језичких модела, као и начин на који се KB кеш формира и користи током ауторегресивног генерисања текста. Кроз анализу релевантних радова обрађене су групе метода које његов трошак смањују: квантизација KB кеша, избацивање и спајање токена, структурна компресија представе кеша, квантизација параметара модела и управљање кешом на нивоу система. Учинак објављених метода је мерен на серверским графичким процесорима, због чега њихово понашање на рачунарима опште намене није довољно испитано. Као резултат студијског истраживачког рада дефинисана је методологија експерименталног поређења формата KB кеша у систему за закључивање vLLM, која обухвата избор модела, мерне оквире за оцену тачности и метрике заузећа меморије, кашњења и пропусне моћи.

3. Опис мастер рада

Мастер рад обухвата 49 страна, са укупно 5 слика, 13 табела, 1 листингом и 27 референци. Рад садржи увод, 4 поглавља и закључак (укупно 6 поглавља), као и списак коришћене литературе. На крају самог рада се налази и додатак од 4 стране. Рад је написан на српском језику.

Прво поглавље је увод у коме су описани предмет и циљ рада. Изложена је мотивација зашто KB кеш утиче на перформансе закључивања великих језичких модела.

У другом поглављу су представљени основни појмови везани за велике језичке моделе. Дат је преглед развоја језичких модела, дефинисан значај малих језичких модела, описана је трансформер архитектура и њени типови, а затим је објашњен механизам пажње,

укључујући скалирану пажњу са скаларним производом, вишеглаву пажњу и груписану пажњу упита.

Треће поглавље се односи на KB кешеве. Објашњен је поступак ауторегресивног генерисања текста и начин на који се кеш при томе формира и користи. Затим су изложени проблеми који настају при раду са дугим секвенцама са становишта меморијске потрошње и ограничења пропусног опсега меморије.

У четвртном поглављу дат је преглед метода за оптимизацију KB кеша, разврстаних према начину на који смањују његов трошак. Разматране су квантизација KB кеша, избацивање и спајање токена, структурна компресија у представи кеша, квантизација параметара модела и методе на нивоу система.

У петом поглављу су приказани измерени резултати заузећа меморије и капацитета KB кеша, тачности на мерилима и брзине закључивања, уз преглед резултата и препоруке за избор конфигурације.

Шесто поглавље је закључак у коме је дат преглед урађеног и предложени су могући правци за даље унапређење и наставак истраживања.

У додатку је дат преглед параметара свих конфигурација које су покретане у склопу спроведених експеримената.

4. Анализа са кључним резултатима

Мастер рад дипл. инж. Жељка Урошевића бави се актуелном проблематиком меморијске потрошње КВ кеша, чије заузеће расте линеарно са дужином контекста и бројем истовремено опслужених захтева. На уређајима са ограниченом меморијом управо КВ кеш одређује да ли је модел могуће покренути и колико захтева се може обрађивати истовремено, па избор његовог формата постаје пројектна одлука од практичног значаја.

У оквиру рада, кандидат је систематизовао постојеће методе за смањење меморијске потрошње КВ кеша и експериментално проверио одабране методе на рачунару опште намене са графичком картицом од 8 GiB меморије. Мерења су спроведена у систему за закључивање vLLM, на три модела и шест конфигурација КВ кеша, при чему су мерени заузеће меморије, капацитет кеша, тачност на стандардизованим мерилима и брзина закључивања.

Мастер рад кандидата садржи више значајних резултата и доприноса, међу којима се издвајају следећи:

- Систематизација метода за смањење трошка КВ кеша у групе према начину на који тај трошак смањују, уз преглед репрезентативних метода и компромиса између постигнутог степена компресије и тачности.
- Постављање поновљивог мерног оквира у систему vLLM, који обухвата осамнаест конфигурација и три групе мерења заузећа меморије, тачности и брзине закључивања.
- Показивање да се измерено заузеће кеша по токenu поклапа са теоријски изведеним вредностима и да фактор компресије зависи од формата кеша, а не од модела.
- Показивање да уштеђена меморија не доноси убрзање сама по себи, већ да компресија повећава пропусну моћ, до 1,52 пута, само када је КВ кеш заиста ограничавајући чинилац.

Резултати рада показују да компресија КВ кеша првенствено повећава број захтева који се могу истовремено обрађивати и да њена корист зависи од модела, архитектуре пажње и расположиве меморије, чиме рад даје практичне смернице за избор конфигурације.

5. Закључак и предлог

Кандидат Жељко Урошевић је у свом мастер раду успешно систематизовао методе за смањење меморијског трошка КВ кеша и спровео опсежну експерименталну проверу одабраних метода на рачунару опште намене, повезујући преглед постојећих решења и сопствена мерења у јединствену целину. Добијени резултати и постављени мерни поступак дају допринос области ефикасног закључивања великих језичких модела.

Кандидат је током израде рада показао висок ниво самосталности и систематичности, као и способност примене теоријских концепата у практичној имплементацији и критичког тумачења добијених резултата. Рад одликује јасна структура, прецизност и систематичност у излагању.

На основу изложеног, Комисија предлаже Комисији за студије II степена Електротехничког факултета у Београду да рад дипл. инж. Жељка Урошевића под насловом „Оптимизација закључивања великих језичких модела кроз квантизацију и компресију КВ кеша” прихвати као мастер рад и кандидату одобри јавну усмену одбрану.

Београд, 30.08.2026. године

Чланови комисије:

др Маја Вукасовић Доцент
сагласан, 29.08.2026.

Никола Станковић Асистент
сагласан, 30.08.2026.