

КОМИСИЈИ ЗА СТУДИЈЕ II СТЕПЕНА ЕЛЕКТРОТЕХНИЧКОГ ФАКУЛТЕТА У БЕОГРАДУ

Комисија за студије II степена, Електротехничког факултета у Београду, на својој седници одржаној 08.09.2026 године именовала нас је у Комисију за преглед и оцену мастер рада дипл. инж. Алекса Ђукић под насловом „Оптимизација ауторегресивне инференције језичких модела заснованих на трансформерској архитектури". Након прегледа материјала Комисија подноси следећи

ИЗВЕШТАЈ

1. Биографски подаци кандидата

Алекса Ђукић, рођен 28.6.2001. године у Београду, студент је мастер академских студија на Електротехничком факултету Универзитета у Београду, на модулу Сигнали и системи, које похађа од 2024. године. На истом модулу похађао је и основне академске студије, у периоду од 2020. до 2024. године. У оквиру основних студија израдио је дипломски рад под насловом „Методе за процену симетрије термографских слика лица за примену у дијагностици“, који је одбранио под менторством проф. др Надице Миљковић.

Пре уписа на факултет похађао је Основну школу „Влада Аксентијевић“ у Београду, коју је завршио 2016. године као вуковац, а потом и Математичку гимназију у Београду, коју је завршио 2020. године.

2. Извештај о студијском истраживачком раду

Кандидат Алекса Ђукић је као припрему за израду мастер рада спровео истраживање релевантне литературе из области оптимизације ауторегресивне инференције језичких модела заснованих на трансформерској архитектури. Посебна пажња посвећена је анализи два приступа који делују на различите изворе неефикасности ауторегресивног генерисања: KV cache механизма и speculative decoding приступа. Анализом литературе размотрена су рачунарска и меморијска ограничења инференције, као и утицај дужине контекста, величине серије и начина извршавања на расположивом хардверу. Додатно, уочено је да се оптимизацији инференције може приступити на више нивоа, укључујући смањење меморијског трошка KV cache-а квантизацијом и оптимизацију механизма пажње кроз ефикаснију организацију приступа GPU меморији. На основу спроведеног истраживања дефинисан је циљ мастер рада, који обухвата реализацију и експерименталну анализу KV cache механизма и greedy speculative decoding алгоритма у оквиру minGPT имплементације GPT архитектуре. Циљ је испитивање утицаја наведених метода на перформансе инференције и утврђивање услова у којима њихова примена може донети мерљиву практичну корист. Полазна minGPT имплементација проширена је тако да омогући засебну примену обе методе и њихову комбинацију, уз задржавање основног начина инференције ради поређења.

3. Опис мастер рада

Мастер рад обухвата 54 стране, са укупно 19 слика, 9 табела и 12 референци. Рад садржи увод, 3 поглавља и закључак, као и списак коришћене литературе, списак скраћеница, списак слика, списак табела.

У уводном поглављу описани су предмет, мотивација и циљ рада. Представљен је проблем оптимизације ауторегресивне инференције језичких модела заснованих на трансформерској архитектури. У другом поглављу представљене су теоријске основе оптимизације ауторегресивне инференције, укључујући основне елементе трансформерске архитектуре и механизма пажње, KV cache механизам и speculative decoding, као и друге приступе оптимизацији инференције. У трећем поглављу описани су експериментално окружење и реализација анализираних метода. Представљени су полазна minGPT имплементација и коришћени GPT-2 модели, реализација KV cache механизма и greedy speculative decoding алгоритма, као и методологија мерења и поступак верификације исправности имплементације. Четврто поглавље садржи резултате експерименталне анализе, која обухвата испитивање утицаја KV cache механизма на време генерисања и меморијску потрошњу, анализу стопе прихватања и утицаја параметра γ на перформансе speculative decoding приступа, као и поређење аналитичког и експерименталног избора параметра γ и четири конфигурације инференције. Пето поглавље представља закључак рада, у оквиру кога су сумирани остварени резултати и сагледан значај примене KV cache механизма и greedy speculative decoding алгоритма у оптимизацији ауторегресивне инференције. Размотрена су ограничења реализованих метода и предложени правци даљег истраживања.

4. Анализа са кључним резултатима

Мастер рад дипл. инж. Алексе Ђукића посвећен је реализацији и експерименталној анализи метода за оптимизацију ауторегресивне инференције језичких модела заснованих на трансформерској архитектури. У оквиру minGPT имплементације реализовани су KV cache механизам и greedy speculative decoding алгоритам, засебно и у комбинацији, а експериментална анализа спроведена је на моделима GPT2 и GPT2-medium. Резултати су показали да KV cache смањује време генерисања избегавањем поновног израчунавања кључева и вредности за већ обрађене позиције, при чему његова ефикасност постаје израженија са повећањем дужине контекста. Анализа speculative decoding приступа показала је да перформансе зависе од стопе прихватања предложених токена, односа трошкова draft и target модела и вредности параметра γ . Највећа просечна пропусност износила је 52,03 токена у секунди, што представља убрзање од 1,32 пута у односу на конфигурацију са KV cache механизмом. На крају је извршена упоредна анализа основне ауторегресивне инференције, KV cache механизма, speculative decoding приступа и њихове комбинације. Основни резултати рада су: 1) имплементација KV cache механизма и анализа утицаја на време генерисања и меморијску потрошњу, 2) имплементација speculative decoding приступа и испитивање утицаја стопе прихватања, и 3) упоредна анализа појединачне и комбиноване примене реализованих метода оптимизације.

5. Закључак и предлог

Кандидат Алекса Ђукић је у свом мастер раду проблем оптимизације ауторегресивне инференције језичких модела заснованих на трансформерској архитектури. У оквиру рада реализовани су и експериментално испитани KV cache механизам и greedy speculative decoding алгоритам, појединачно и у комбинацији, у оквиру minGPT имплементације и на моделима GPT-2. Анализиран је њихов утицај на време генерисања, меморијску потрошњу и пропусност, као и услови у којима ове методе могу допринети ефикаснијој инференцији. 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. Алексе Ђукића прихвати као мастер рад и кандидату одобри јавну усмену одбрану.

Београд, 12.09.2026. године

Чланови комисије:

др Марија Новичић Доцент
сагласан, 12.09.2026.

Наталија Ђорђевић Асистент
сагласан, 12.09.2026.