

КОМИСИЈИ ЗА СТУДИЈЕ II СТЕПЕНА ЕЛЕКТРОТЕХНИЧКОГ ФАКУЛТЕТА У БЕОГРАДУ

Комисија за студије II степена, Електротехничког факултета у Београду, на својој седници одржаној 25.08.2026 године именовала нас је у Комисију за преглед и оцену мастер рада дипл. инж. Андреј Ачковић под насловом „Оптимизација извршавања Diffusion Transformer модела за генерисање видео записа“. Након прегледа материјала Комисија подноси следећи

ИЗВЕШТАЈ

1. Биографски подаци кандидата

Андреј Ачковић је рођен 25.12.2002. године у Нишу. Завршио је основну школу „Душан Радовић“ у Нишу као вуковац и ђак генерације, а затим похађао гимназију „Бора Станковић“ у Нишу коју завршава 2021. године такође као вуковац. У октобру исте године уписује Електротехнички факултет у Београду, где се одлучује за смер „Софтверско инжењерство“. Дипломирао је на истом одсеку 2025. године. Дипломски рад на тему „Систем за евиденцију завршних радова заснован на MEAN архитектури“ одбранио је октобра 2025. године са оценом 10, код ментора доц. др Владимира Јоцковића. Након тога у новембру исте године уписује мастер студије на Електротехничком факултету у Београду, на модулу за Софтверско инжењерство.

2. Извештај о студијском истраживачком раду

Кандидат Андреј Ачковић је као припрему за израду мастер рада урадио истраживање литературе у областима генерисања видеа записа и трансформер архитектура, као и наменског хардвера за обучавање модела машинског учења и закључивање, као што је Tenstorrent архитектура. Анализирани су дифузиони модели, као што су Stable Diffusion и HunyuanVideo, као и њихова примена у генерисању видеа. Уочен је значајан изазов у обради великог броја токена и повећано рачунарско оптерећење у процесу генерисања. Истраживањем су идентификовани проблеми повезани са дистрибуираним извршавањем и оптимизацијом механизма пажње. На основу тога дефинисан је концепт решења које користи TT-XLA и TT-MLIR за оптимизацију извршавања модела Wan 2.2 T2V-A14B на Tenstorrent хардверу, што је довело до значајног подобраја перформанси и смањења захтева за рачунарским ресурсима.

3. Опис мастер рада

Мастер рад има 60 страница, укључујући 13 слика, 3 табеле и 31 референци. Рад садржи увод, 4 поглавља и закључак (укупно 6 поглавља), списак коришћене литературе, скраћеница, слика и табела. Рад је написан на српском језику.

У другом поглављу анализирају се модели за генерисање видео записа и трансформер архитектура, као и њихова примена у процесу дифузије. Описује се развој дифузионих модела, укључујући Stable Diffusion, HunyuanVideo и Diffusion Transformer, као и архитектура модела Wan 2.2 T2V-A14B.

У трећем поглављу представљено је Tenstorrent архитектура и њена организација, укључујући Tensix језгра, меморијску структуру, комуникациону инфраструктуру и софтверски стек за ефикасно извршавање модела машинског учења. Описан је процес превођења модела кроз различите нивое апстракције, од тензорских операција до директног извршавања на хардверу.

У оквиру четвртог поглавља приказан је процес оптимизације извршавања модела Wan 2.2 T2V-A14B кроз компајлерску путању TT-XLA и TT-MLIR, укључујући идентификацију разлика у перформансама у поређењу са референтном имплементацијом и увод трансформација које смањују ове разлике. Описују се конкретне оптимизације као што су спајање условљавања и нормализације, ротационо кодирање, обједињавање колективних операција са матричним множењем, а такође прстенаста самопажња и расподељена редукција.

Пето поглавље доноси анализу перформанси компајлерског графа у поређењу са референтном имплементацијом, укључујући мерења времена извршавања и броја операција за различите оптимизације и топологије. Описује како свака оптимизација утиче на брзину и ефикасност, као и преостали јаз између компајлерског и референтног извршавања. Шесто поглавље приказује закључак рада са правцима даљег истраживања.

4. Анализа са кључним резултатима

У раду је анализирана оптимизација извршавања Diffusion Transformer модела за генерисање видеозаписа на специјализованом Tenstorrent хардверу, са акцентом на компајлерске методе и њихов утицај на перформансе. У раду је показано да се део губитка перформанси при извршавању великог генеративног модела може отклонити на нивоу компајлера, без измене самог модела и без ручног писања кернела. Основни допринос рада је развој компајлерских оптимизација кроз MLIR и TT-MLIR, које препознају и оптимизују рачунске и комуникационе обрасце у међурепрезентацији, што доводи до значајног побољшања перформанси у поређењу са почетном путањом кроз TT-XLA. Добитак од ових оптимизација је систематски фокусиран увођењем пролаза који препознају обрасце у графу, што омогућава ефикасну и уопштенију примену. Основни резултати рада су: 1) ефикасна компајлерска путања за оптимизацију извршавања модела, 2) анализа и поређење перформанси пре и након оптимизација, 3) утврђивање да компајлерске оптимизације могу да се примене на широк спектар блокова и процеса, а не само на конкретан модел.

5. Закључак и предлог

Кандидат Андреј Ачковић је у свом мастер раду успешно показао како се део губитка перформанси при извршавању великог генеративног модела за видео на специјализованом хардверу може отклонити на нивоу компајлера, без измене самог модела и без ручног писања кернела. Развијене су компајлерске оптимизације кроз MLIR и TT-MLIR, које су довеле до значајног побољшања перформанси у поређењу са почетном путањом кроз TT-XLA.

На основу изложеног, Комисија предлаже Комисији за студије II степена Електротехничког факултета у Београду да рад дипл. инж. Андреја Ачковића под насловом „Оптимизација извршавања Diffusion Transformer модела за генерисање видеозаписа“ прихвати као мастер рад и кандидату одобри јавну усмену одбрану.

Београд, 30.08.2026. године

Чланови комисије:

др Марко Мишић Ванредни професор
сагласан, 29.08.2026.

др Драгомир Ел Мезени Доцент
сагласан, 30.08.2026.

Матија Додовић Асистент
сагласан, 29.08.2026.