

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. Филип Ескић под насловом „TONIC: Преводац ONNX модела неуралних мрежа у C/C++ код заснован на MLIR радном оквиру". Након прегледа материјала Комисија подноси следећи

ИЗВЕШТАЈ

1. Биографски подаци кандидата

Филип Ескић је рођен 11. децембра 1999. године у Београду. Основну школу Живомир Савковић у Ковачевцу завршио је као носилац Вукове дипломе. Након тога је уписао Техничку школу у Младеновцу, смер Електротехничар информатичких технологија, коју је завршио као носилац Вукове дипломе и најбољи ученик на смеру.

Основне академске студије на Електротехничком факултету Универзитета у Београду уписао је 2018. године. Студије је започео на одсеку за Телекомуникације и информационе технологије, након чега је променио студијски програм и наставио студије на одсеку за Рачунарску технику и информатику. Дипломски рад под називом „Унапређивање модела за детекцију бициклиста у саобраћају коришћењем техника дубоког учења“ одбранио је у септембру 2024. године са оценом 10, под менторством професора др Марка Мишића.

Мастер академске студије на Електротехничком факултету Универзитета у Београду уписао је 2024. године.

Професионалну каријеру започео је стручном праксом у компанији TTTech Auto, где је након завршетка праксе наставио рад на позицији софтверског инжењера. Тренутно је запослен у компанији TrustMotion, где се бави развојем софтвера у области Edge AI система.

2. Извештај о студијском истраживачком раду

Кандидат Филип Ескић је као припрему за израду мастер рада урадио истраживање релевантне литературе која се односи на област компајлерских радних оквира за машинско учење, којој припада тема мастер рада. Истраживање је разматрало случај употребе где се фиксан, унапред познат модел неуралне мреже специјализује у време превожња и представља самосталним C/C++ кодом. Истраживањем области кандидат је утврдио да је MLIR инфраструктура погодан приступ за решавање овог типа проблема. Проучена је MLIR инфраструктура, начин на који се у њој формира међурепрезентација, методологија прогресивног спуштања, а такође и сам ONNX формат за представљање конкретних модела. Истраживањем је утврђено да постојећа решења (ONNX-MLIR, MLIR-EmitC, emx-onnx-cgen) полазе од другог улазног формата, као крајњи производ дају објектни фајл и библиотеку које корисник тек треба да повеже или уопште не користе MLIR инфраструктуру. Одатле, случај у коме се, кроз MLIR инфраструктуру, из ONNX модела добија самосталан C/C++ код без извршног окружења и динамичке алокације меморије, остаје непокривен.

3. Опис мастер рада

Мастер рад обухвата 71 страну, са укупно 7 слика, 6 табела, 49 листинга програмског кода и 69 референци. На крају рада се налази додаток на 20 страна. Рад садржи увод, пет поглавља и закључак (укупно седам поглавља), списак коришћене литературе, списак скраћеница, слика, табела и листинга, као и прилог. Прво поглавље представља увод, у коме су описани предмет и циљ рада. Разматра се потреба за извршавањем модела машинског учења на edge уређајима и формира полазна претпоставка рада: за унапред познат и фиксан модел, израчунавање се може специјализовати у време превожња и добити C/C++ код са уграђеним тежинама. Из тако добијеног кода, даљим превожњем настаје извршни програм, који не захтева извршно окружење ни динамичку алокацију меморије. У другом поглављу детаљно је представљена MLIR инфраструктура. Треће поглавље даје преглед сродних решења, груписаних према радним оквирима

заснованим на MLIR инфраструктури, приступима генерисању C/C++ кода и алатима за превођење модела на уграђене системе. Четврто поглавље описује пројектовање система. Дати су циљеви преводиоца и ограничења која из њих произилазе, архитектура тока превођења, наменски дијалекат `tonic`, као и кључне пројектне одлуке. Пето поглавље се бави имплементацијом. Описани су парсер `onnx2tonic`, дефиниција `tonic` дијалекта и `tonic-opt` алат, верификатори, коришћени интерфејси и пролаз спуштања. Поглавље обухвата и измену постојећег MLIR кода као и приказ генерисаног кода и пратећих фајлова. Шесто поглавље представља валидацију решења. Описани су одабрани модели, скупови података и методологија рада, а затим су изложени и анализирани добијени резултати. Седмо поглавље је закључак, у коме су резимирани приступ и добијени резултати и дате смернице за даље унапређење система. Рад садржи и прилог, у коме су детаљније образложене поједине пројектне одлуке, приказано праћење једног примера кроз све нивое превода и дато упутство за понављање поступка валидације.

4. Анализа са кључним резултатима

Мастер рад дипл. инж. Филипа Ескића се бави превођењем истренираног модела неуралне мреже ONNX формата у C/C++ код, користећи MLIR радни оквир. За те потребе имплементиран је TONIC преводилац, који кроз наменски дијалекат `tonic` и седам подржаних ONNX оператора методологијом прогресивног спуштања до дијалекта `emitc` формира изворни код на језику C/C++. Добијени код има уграђене све тежине разматраних модела и не користи динамичку алокацију меморије.

Исправност приступа испитана је поређењем са референтним алатом ONNX Runtime, над три модела растуће сложености. Разматрано је укупно 20150 узорака, где је постигнуто потпуно слагање предвиђених класа, а највеће забележено нумеричко одступање износи $2,5 \cdot 10^{-6}$.

Основни доприноси рада су:

1. Имплементација наменског MLIR дијалекта и комплетног тока превођења с краја на крај.
2. Експериментална потврда добијеног решења поређењем са референтним алатом ONNX Runtime.
3. Могућност наставка рада, кроз проширење скупа подржаних оператора и увођење оптимизација специфичних за машинско учење.

5. Закључак и предлог

Кандидат Филип Ескић је у свом мастер раду успешно решио проблем превођења статички познатог модела неуралне мреже ONNX формата у самосталан код на језику C/C++, применом методологије прогресивног спуштања кроз MLIR инфраструктуру. Развијено решење генерише код који у себи садржи параметре модела и не користи динамичку алокацију меморије. Исправност решења потврђена је поређењем са референтним алатом ONNX Runtime, при чему је на свим посматраним узорцима остварено потпуно слагање предвиђених класа, уз нумеричка одступања која су у складу са очекиваним последицама заокруживања подржаног типа float32.

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. Филипа Ескића прихвати као мастер рад и кандидату одобри јавну усмену одбрану.

Београд, 11.09.2026. године

Чланови комисије:

др Маја Вукасовић Доцент
сагласан, 11.09.2026.

Никола Станковић Асистент
сагласан, 11.09.2026.