# Programski prevodioci za mašinsko učenje

**CIlj kursa:**

Upoznavanje sa:

- procesom prevođenja modela mašinskog učenja,
- grafovima izračunavanja i kako predstaviti model preko grafa,
- osnovnim tehnikama i koracima za transformaciju graf modela u optimizovanu verziju za odgovarajući hardver.

Praktične vežbe koriste TVM [1] i njegov eko-sistem.

**Očekivana predznanja:**

Poznavanje osnovnih pojmova iz operativnih sistema, kao i osnove programiranja u C-u. Poželjno poznavanje Pythona i osnova programskih prevodilaca.

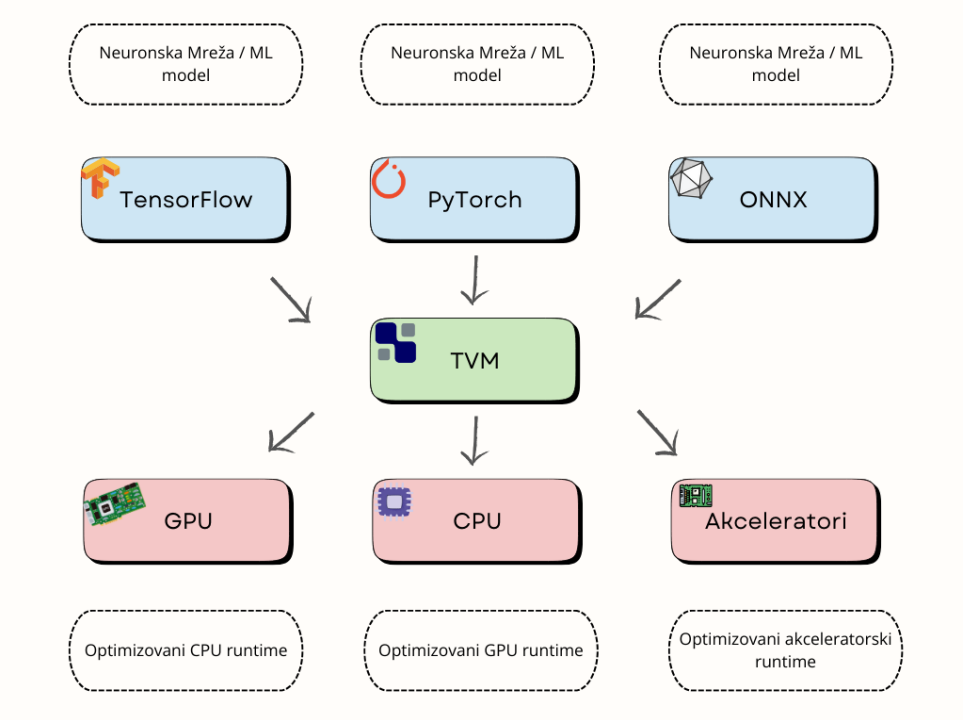
**Tehnologije:**

C, C++, Python, PyTorch, TVM, LLVM/MLIR, Linux.

**Sadržaj kursa**:

Mašinsko učenje je poslednjih godina postalo veoma popularno. Usled ovoga je porasla potreba da se modeli mašinskog učenja izvršavaju u sve raznovrsnijim kontekstima i na sve različitijem hardveru. Danas je najčešće rešenje za ovaj izazov korišćenje namenskih programskih prevodilaca, pomoću kojih se jedan model može prevesti na više različitih hardverskih jedinica. Cilj ovog kursa je da se upoznamo sa datim tehnologijama, vidimo kako se koriste i kako se mogu unaprediti.

Početni deo kursa ima za cilj da studente upozna sa osnovnim konceptima u ovoj oblasti kao i kako prevesti modele mašinskog učenja. Da bi preveli modele, studenti će koristiti TVM koji je već nekoliko godina najcelovitiji i najpopularniji ekosistem za programske prevodioce za mašinsko učenje koji koriste velike svetske kompanije (Amazon, AMD, ARM, Facebook, Intel, Microsoft, Qualcomm). Za modele ćemo uzeti primere napisane koristeći radni okvir PyTorch [2].




Drugi deo kursa se fokusira na reprezentaciju modela i optimizacijama koje se nad njima primenjuju. Biće dat pregled najčešćih operacija koje se koriste u aktuelnim modelima kao i grafovska reprezentacija toka izračunavanja. U ovom delu kursa, studenti će se upoznati sa međukodom Relay, formatom za grafovsku reprezentaciju neuronskih mreža. Pored toga, biće reči i o optimizacijama na nivou grafa, kao što su izračunavanje konstanti i fuzija operatora.

U trećem delu se obrađuje reprezentacija grafova nižeg nivoa. Studenti će naučiti više o apstrakciji tenzorskih programa i primitiva tenzorskih izračunavanja. Nakon toga, biće prikazane optimizacije na nivou tenzorskih operacija kao što su blokovski pristup susednim podacima, vektorizacija, permutacija petlji, pakovanje nizova za efikasniji pristup memoriji.

Pretposlednji deo kursa se bavi finim podešavanjem modela i njihovom optimizacijom prilikom prevođenja. Nakon toga, detaljnije se analizira prevođenje mašinskog koda modela, i koje su razlike ukoliko se model prevodi za GPU, CPU ili neki akcelerator. Studenti će se upoznati sa reprezentacijom modela ONNX [3] koju TVM koristi kao osnovnu, a pokriveno je i prilagođavanje modela iz PyTorch. Konačno, aspekti raspoređivanja prevedenih modela i njihova upotreba unutar drugih aplikacija će biti prikazani. Studenti će imati mogućnost da u okviru praktičnog dela steknu iskustva u korišćenju alata za prevođenje, i da ih primene na modele otvorenog koda za klasifikaciju slika kao i velike jezičke modele.

Poslednji deo kursa obuhvata i druge relevantne tehnologije, sa akcentom na radni okvir Torch-MLIR [4][5]. Kao praktičan deo, biće obuhvaćena realizacija novog operatora sa dekompozicijom na već postojeće.

**Literatura**:

[1] https://tvm.apache.org/
[2] https://pytorch.org/
[3] https://onnx.ai/
[4] https://mlir.llvm.org/
[5] https://github.com/llvm/torch-mlir

Studentima će se takođe nakon svakog časa dostaviti literatura vezana za prezentovanu materiju.