Univerzitet Crne Gore
Prirodno-matematički fakultet

Džordža Vašingtona b.b.
1000 Podgorica, Crna Gora

tel: +382 (0)20 245 204
fax: +382 (0)20 245 204
www.pmf.ac.me

Broj: 2023/01-1977

Datum: 25. 09. 2023

Vijeću Prirodno-matematičkog fakulteta

Molim Vijeće Prirodno-matematičkog fakulteta da odobri uvođenje izbornog predmeta *Optimizacija i kompresija modela dubokog učenja* na doktorskim studijama računarskih nauka.

Optimizacija i kompresija modela dubokog učenja je nova disciplina koja dobija na važnosti zbog sve većih modela dubokog učenja ali i i aktuelne tendencije za dizajnom što bržih algoritama sa što manjim hardverskim zahtjevima, te i manjom potrošnjom energije. Znanja iz ove oblasti su neophodna da bi modeli dubokog mašinskog učenja bili upotrebljivi u većini sektora i približavaju doktorske studije našeg fakulteta potrebama moderne industrije.

Ciljevi kursa su pregled tehnika za traženje optimalne arhitekture i kompresiju neuronskih mreža, korišćenje postojećih aktuelnih tehnologija i radnih okvira, uz praktičan rad sa edge uređajima koje Prirodno-matematički fakultet već posjeduje zahvaljujući partnerima iz industrije.

U Podgorici
25.09.2023.

Doc. dr Igor Jovančević

Igor Jovančević

Tabela S2.6.4. Forma za pripremu informacionih lista predmeta

Naziv predmeta	Optimizacija i kompresija modela dubokog učenja			
Šifra predmeta	Status predmeta	Semestar	Broj ECTS kredita	Fond časova
	izborni	II	5	4P+0V

Studijski programi za koje se organizuje
Računarske nauke (akademske doktorske studije, studije traju 6 semestara, 180 ECTS kredita)

Uslovljenost drugim predmetima

Ciljevi izučavanja predmeta
Izučavanjem ovog predmeta studenti se upoznaju sa savremenim metodama optimizacije i kompresije modela dubokog učenja, sa ciljem njihovog bržeg izvršavanja i smanjenja hardverskog opterećenja. Obrađuju se tehnike kao što je traženje optimalne arhitekture neuronske mreže (Neural Architecture Search - NAS), destilacija znanja (Knowledge Distillation - KD), kvantizacija i pruning neuronskih mreža. Takođe, izučavanje ovog predmeta obuhvata savladavanje nekih od već dostupnih tehnologija za optimizaciju (TFMOT, TensorFlow Lite, TensorRT, Optuna), kao i rad sa *edge* uređajima, prvenstveno iz NVidia Jetson familije.

Sadržaj predmeta (nastavne cjeline, oblici individualnog rada studenata, oblici provjere znanja) prikazan prema radnim nedjeljama u akademskom kalendaru:

Pripremna nedjelja	
I nedjelja	Duboko učenje (Deep Learning - DL) i neke vrste neuronskih mreža: kovolucione, mreže za obradu 3D podataka, transformeri. Brzina izvršavanja i hardverski zahtjevi modela dubokog učenja, potreba za njihovom optimizacijom i kompresijom. Edge uređaji i izvršavanje modela na njima. Pregled TensorFlow biblioteke za duboko učenje u Python-u.
II nedjelja	Pregled najčešće korišćenih tehnika za optimizaciju dubokih neuronskih mreža: kvantizacija, pruning, NAS (Neural Architecture Search), destilacija znanja (knowledge distillation). Pregled mogućnosti TFMOT biblioteke za optimizaciju TensorFlow modela. TensorFlow Lite format za efikasnije izvršavanje modela.
III nedjelja	Kvantizacija neuronskih mreža. Prelaz sa floating-point operacija na cjelobrojne operacije sa manjim brojem bita. Varijante: PTQ (Post-training quantization), QAT (Quantization-Aware Training). Kalibracioni skup podataka. ONNX (Open Neural Network eXchange) kao univerzalni format za sladištenje modela dubokog učenja. ONNX-Runtime biblioteka za izvršavanje modela u ONNX formatu. ONNX kvantizacija.
IV nedjelja	Pruning: odbacivanje djelova mreže koji nemaju značajan uticaj na krajnji izlaz. Primjena pruning funkcionalnosti iz TFMOT biblioteke za optimizaciju neke duboke neuronske mreže.
V nedjelja	Automatizovano traženje optimalne arhitekture neuronske mreže za određeni zadatak (NAS - Neural Architecture Search). Najčešće korišćene strategije za pretragu: reinforcement learning i evolucione strategije.
VI nedjelja	Drugi poznati algoritmi za optimizaciju (simulated annealing, ant colony, particle swarm i sl.). Njihove mogućnosti i ograničenja u NAS primjenama. Optuna framework kao kombinacija više poznatih optimizacionih algoritama i heuristika i mogućnost njegovog korišćenja za implementaciju NAS tehnika.
VII nedjelja	Parametrizacija arhitekture neuronske mreže za primjenu NAS tehnika. Prostor pretrage. Parametrizacija konvolucionih neuronskih mreža, optimizacija broja slojeva, broja filtera unutar svakog sloja i veličine konvolucionih kernela. Implementacija NAS pretrage za neku konvolucionu mrežu pomoću Optuna framework-a.
VIII nedjelja	Destilacija znanja (knowledge distillation - KD). Teacher-student pristup. Prenos znanja sadržanog u većoj neuronskoj mreži (učitelj) na manju mrežu (učenik). Formiranje funkcije gubitka za obučavanje mreža KD pristupom. Implementacija KD pristupa za optimizaciju nekog modela dubokog učenja.
IX nedjelja	Kombinovanje NAS i KD tehnika. Traženje optimalne arhitekture student mreže. Implementacija ovakvog pristupa.
X nedjelja	Optimizacije niskog nivoa. Spajanje više slojeva neuronske mreže u jednu operaciju (layer fusion). Uticaj implementacije elementarnih operacija na performanse. Zamjena običnih konvolucija operacijom konvolucije separabilne po dubini (depthwise separable convolution). TensorRT - optimizacija modela za izvršavanje na NVidia grafičkim karticama.
XI nedjelja	Izvršavanje modela dubokog učenja na uređajima malog kapaciteta (edge devices). Minimizacija zahtjeva u pogledu RAM i VRAM memorije. Hardware-in-the-loop optimizacija. Pokretanje nekog modela na jednom od dostupnih edge uređaja (Nvidia Jetson uređaji, Raspberry Pi). Implementacija hardware-in-the-loop procedure za optimizaciju.
XII nedjelja	Posebni hardverski akceleratori - NVidia DLA (Deep Learning Accelerator), TPU (Tensor Processing Unit) i slično. TensorRT optimizacija modela sa DLA podrškom na uređaju Jetson Orin AGX. Poređenje performansi sa i bez DLA podrške.

XIII nedjelja	Optimizacija paralelnog izvršavanja više modela na jednom uređaju. Balansiranje hardverskih resursa između modela. Automatizacija ovakve procedure.
XIV nedjelja	Optimizacija energetske efikasnosti u izvršavanju modela dubokog učenja i njena važnost za budućnost sistema vještačke inteligencije. Razlika između izvršavanja modela na serverskim grafičkim karticama (snage od nekoliko desetina do nekoliko stotina vati) i grafičkim karticama manje snanje (nekoliko vati do nekoliko desetina vati - NVidia Jetson uređaji).
XV nedjelja	Odbrane projekata.

Metode obrazovanja
Predavanja, praćenje rada studenata na praktičnim projektima.

Opterećenje studenata

Nedjeljno	U semestru
5 x 40/30 = 6 sati i 40 minuta Predavanja: 4 sata Vježbe: 0 sati Ostale nastavne aktivnosti: 0 Individualni rad studenata: 2 sata i 40 minuta	**Nastava i završni ispit: (6 sati i 40 minuta) x 16 = 106 sati i 40 minuta** **Neophodne pripreme** (administracija, upis, ovjera prije početka semestra): 2 x (**6 sati i 40 minuta**) = **13 sati i 20 minuta** **Ukupno opterećenje za predmet : 5x30 = 150 sati** **Dopunski rad: od 0 do 30 sati** **Struktura opterećenja**: 106 sati i 40 min(Nastava) + 13 sati i 20 minuta (Priprema) + 30 sati (Dopunski rad)

Obaveze studenata u toku nastave:
Prisustvo nastavi, rad na praktičnom projektu, pisanje seminarskog rada i polaganje završnog ispita.

Literatura:

1. **Hutter, F., Kotthoff, L., & Vanschoren, J. (2019). Automated Machine Learning. Springer.**
2. **Sun, Y., Yen, G., & Zhang, M. (2023). Evolutionary Deep Neural Architecture Search: Fundamentals, Methods, and Recent Advances. Springer.**
3. **Pedrycz, W., & Chen, S. (2023). Advancements in Knowledge Distillation: Towards New Horizons of Intelligent Systems. Springer.**
4. **Gridin, I. (2022). Automated Deep Learning Using Neural Network Intelligence: Develop and Design PyTorch and TensorFlow Models Using Python. O'Reilly.**
1. **Naučni radovi navedeni tokom predavanja**

Ishodi učenja (usklađeni sa ishodima za studijski program):
Student se obučava da optimizuje modele dubokog učenja tako da se smanji njihovo vrijeme izvršavanja, zahtjevi u pogledu RAM/VRAM memorije i slično, a da se pri tome zadrži zadovoljavajuća preciznost u rješavanju nekog problema. Nakon što položi ovaj ispit, student će imati kompetencije da:

1. **Vlada konceptima kao što su traženje optimalne arhitekture neuronske mreže (NAS), destilacija znanja, kvantizacija i pruning neuronskih mreža**
2. **Implementira razne tehnike optimizacije i vlada već dostupnim tehnologijama (Optuna, TFMOT, TensorRT, TensorFlow Lite)**
3. **Razumije važnost optimizacije i kompresije modela (npr. radi izvršavanja na edge uređajima malog kapaciteta)**

Oblici provjere znanja i ocjenjivanje:
Seminarski rad 30 bodova, izrada i odbrana praktičnog projekta 40 bodova i završni ispit 30 bodova. Za prelaznu ocjenu potrebno je imati 51 i više bodova.

Ime i prezime nastavnika: Doc. dr Igor Jovančević I. Jovančević

Napomena (ukoliko je potrebno):