Могућност коришћења националне AI платформе

Поштоване колегинице и колеге,

обавештавам вас да је факултет потписао споразум о сарадњи са Националном платформом за вештачку интелигенцију (Канцеларија за ИТ и еУправу, Државни дата центар у Крагујевцу) и да нам је одобрен приступ суперрачунару те платформе. Приступ је у функцији и већ се користи за истраживања у области дубоког учења на факултету.

Шта је доступно

- GPU кластер заснован на NVIDIA DGX A100 системима (A100 са 40 GB меморије по картици, укупно 32 GPU); у току је проширење системима DGX H200.
- Радно окружење: Linux, Docker контејнери са PyTorch/TensorFlow, JupyterLab на GPU и SLURM ред за дуже послове. Приступ иде преко VPN-а и веб интерфејса (FastML Engine).
- Намена: тренирање и евалуација модела машинског и дубоког учења, обрада великих скупова података и други рачунски захтевни послови у научне и наставне сврхе.

Ко може да користи

- Сви наставници и сарадници факултета, самостално или са студентима основних, мастер и докторских студија: за истраживачке пројекте, докторске дисертације, мастер и завршне радове.
- Сав приступ платформи иде преко асистента Далибора Шељмешија, који поставља код и податке на кластер, покреће послове и враћа резултате. Није потребно да сами отварате налоге нити да учите рад са кластером.

Шта треба припремити

- Код у облику Python скрипти или Jupyter notebook-а који се извршава без интеракције (без ручног уноса током рада), са јасно дефинисаним улазним подацима и местом где се чувају резултати (модели, логови, излазне датотеке).
- Податке у стандардном облику (CSV, слике, NumPy низови и сл.) и списак потребних библиотека.
- Пожељно: чување међурезултата (checkpoint), да би посао могао да се настави ако буде прекинут.

Обавезан први корак: проба на Google Colab-у или Kaggle-у

Пре него што се било шта пошаље на суперрачунар, молим да се експеримент прво покрене на бесплатном GPU-у на Google Colab-у или Kaggle-у (оба нуде NVIDIA T4). Тиме се проверава, прво, да код ради од почетка до краја и да су подаци исправни, и друго, **да ли је посао уопште толико захтеван да му је потребан суперрачунар.**

Ако се цео експеримент на Colab-у или Kaggle-у заврши за сат-два, суперрачунар није потребан и посао остаје тамо. Суперрачунар има смисла за тренирања која на T4 трају данима, за велике скупове података, за моделе који не стају у 16 GB GPU меморије, или за већи број поновљених експеримената (аблације, претрага хиперпараметара). Циљ је да се заједнички ресурс не троши на послове који могу да се заврше на лаптопу или у бесплатном окружењу.

Процедура

1. Прототип на Colab-у или Kaggle-у (линк ка notebook-у или репозиторијуму).
2. Кратак опис мејлом (dalibor.seljmesi@tfzr.rs): циљ експеримента, подаци (врста и величина), процена трајања и броја покретања, носилац (наставник) и студенти који учествују.
3. Заједничко прилагођавање кода за кластер и договор о термину.
4. Покретање на платформи и испорука резултата.

Напомене

- Платформа је намењена искључиво научним и наставним активностима факултета.
- Лични и осетљиви подаци не шаљу се на платформу без одговарајуће сагласности и анонимизације.
- Коришћење платформе обавезно се наводи у свим радовима насталим уз њену помоћ, а о сваком покренутом послу води се евиденција; оба су описана у наставку.

Захвалница у радовима

Коришћење платформе наводи се у сваком раду, саопштењу, мастер раду и дисертацији у којима су резултати добијени на њој, у одељку Захвалница (Acknowledgments). Предлог формулације:

- **Српски:** *Резултати приказани у овом раду добијени су коришћењем рачунарских ресурса Националне платформе за вештачку интелигенцију Републике Србије, смештене у Државном дата центру у Крагујевцу, коју обезбеђује Канцеларија за информационе технологије и електронску управу.*
- **Енглески:** *The results presented in this paper were obtained using the computational resources of the National Artificial Intelligence Platform of the Republic of Serbia, hosted at the State Data Centre in Kragujevac and provided by the Office for Information Technologies and eGovernment.*

Ако рад садржи и одељак о коришћеној опреми (Computational resources, Hardware, Implementation details), у њему треба навести конфигурацију на којој је посао извршен: NVIDIA DGX A100, графички процесори A100-SXM4-40GB, уз приближан број GPU часова по експерименту. Те податке добијате од Далибора Шељмешија заједно са резултатима.

Молим да Далибору Шељмешију текст захвалнице пошаљете пре предаје рада, да би се проверило да ли је формулација у складу са споразумом и да је наведена конфигурација тачна. Ако платформа у међувремену пропише своју обавезну формулацију, овај предлог се замењује њоме.

Евиденција коришћења

О сваком послу покренутом на платформи Далибор Шељмеши води евиденцију. Она је потребна из три разлога: факултет по споразуму извештава о коришћењу додељених ресурса, интерно распоређујемо време када има више захтева него слободних термина и на основу оствареног коришћења образлажемо будуће захтеве за веће квоте. Евиденција за сваки посао садржи:

- датум захтева и датум покретања посла;
- носиоца посла (наставник или сарадник) и студенте који учествују, са нивоом студија;

- назив и кратак опис експеримента, као и пројекат, мастер рад или дисертацију у оквиру којих се ради;
- скуп података (врста и величина) и коришћени модел или библиотеку;
- број заузетих графичких процесора и трајање посла, односно остварене GPU часове;
- исход (завршено, прекинуто, поновљено) и кратку напомену о уоченим проблемима;
- референцу рада или тезе у којој су резултати објављени, када до објаве дође.

Техничке спецификације: https://www.ai.gov.rs/tekst/sr/189/nacionalna-ai-platforma.php