



ПРИРОДНО-МАТЕМАТИЧКИ ФАКУЛТЕТ
УНИВЕРЗИТЕТ У КРАГУЈЕВЦУ

Istaživanje

Predmet

Uvod u softversko inženjerstvo

Tema:

Istraživanje za čuvanje podataka ANN

Student :
Ivan Lubisavljević

Zadatak:

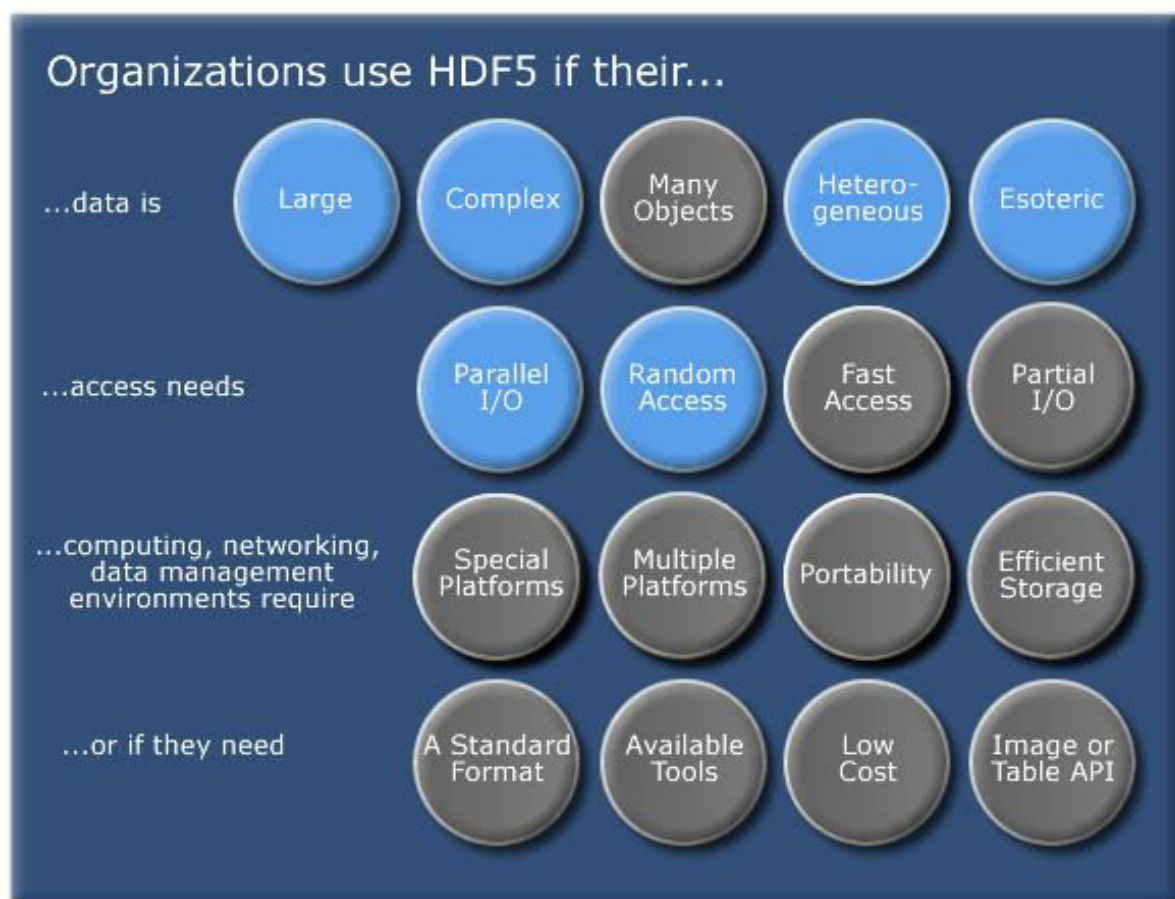
Istražiti HDF5, Onnx formate za skladištenje podataka i napisati kratak izveštaj sa pronađenim prednostima i manama.

Ukratko o formatu H5

Šta je H5 format datoteke?

H5 datoteka je datoteka s podacima sačuvana u Hijerarhijskom formatu podataka (HDF - Hierarchical Data Format). Sadrži višedimenzionalne nizove naučnih podataka. H5 fajlovi se obično koriste u vazduhoplovstvu, fizici, inženjerstvu, finansijama, akademskim istraživanjima, astronomiji, elektronskim instrumentima i medicinskim poljima.

Kada se koristi?



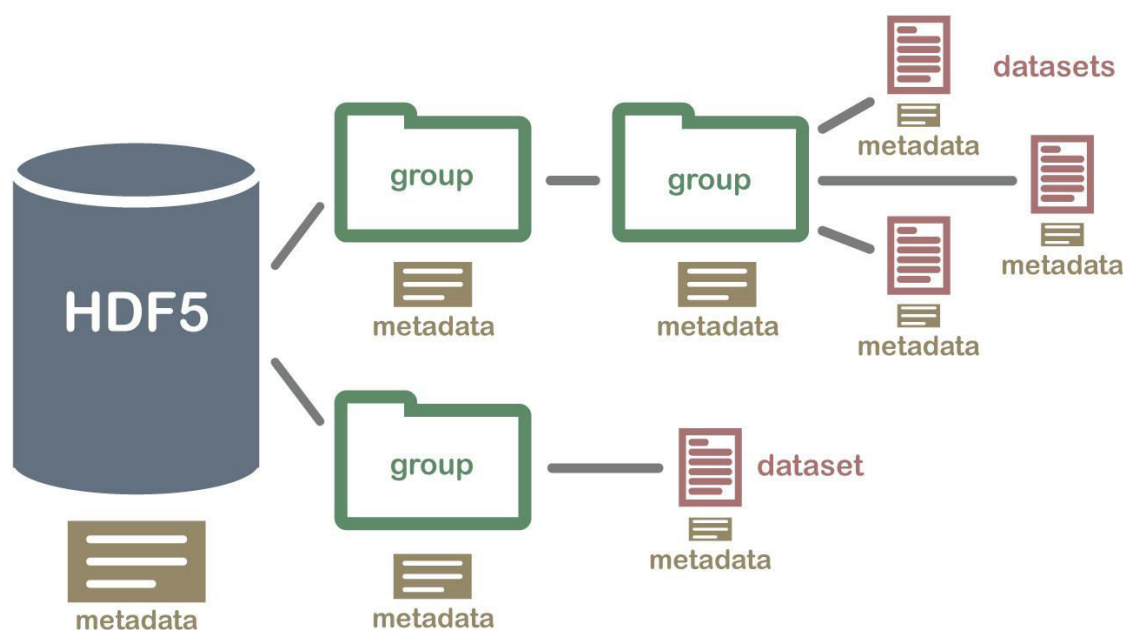
Hijerarhijska struktura – direktorijum sa fajlovima unutar fajla

Format HDF5 se može zamisliti kao sistem datoteka sadržan i opisan unutar jedne datoteke. Razmislite o datotekama i fasciklama koje se nalaze na vašem računaru. Možda imate direktorijum podataka sa nekim podacima o temperaturi za više lokacija na terenu. Ovi podaci o temperaturi se prikupljaju svakog minuta i sumiraju na satnoj, dnevnoj i nedeljnoj bazi. Unutar jedne HDF5 datoteke možete da skladištite sličan skup podataka organizovan na isti način na koji biste mogli da organizujete datoteke i fascikle na svom računaru. Međutim, u HDF5 datoteci, ono što nazivamo "direktorijumi" ili "fascikle" na našim računarima se nazivaju grupe, a ono što nazivamo datotekama na našem računaru se nazivaju skupovi podataka.

2 važna HDF5 uslova

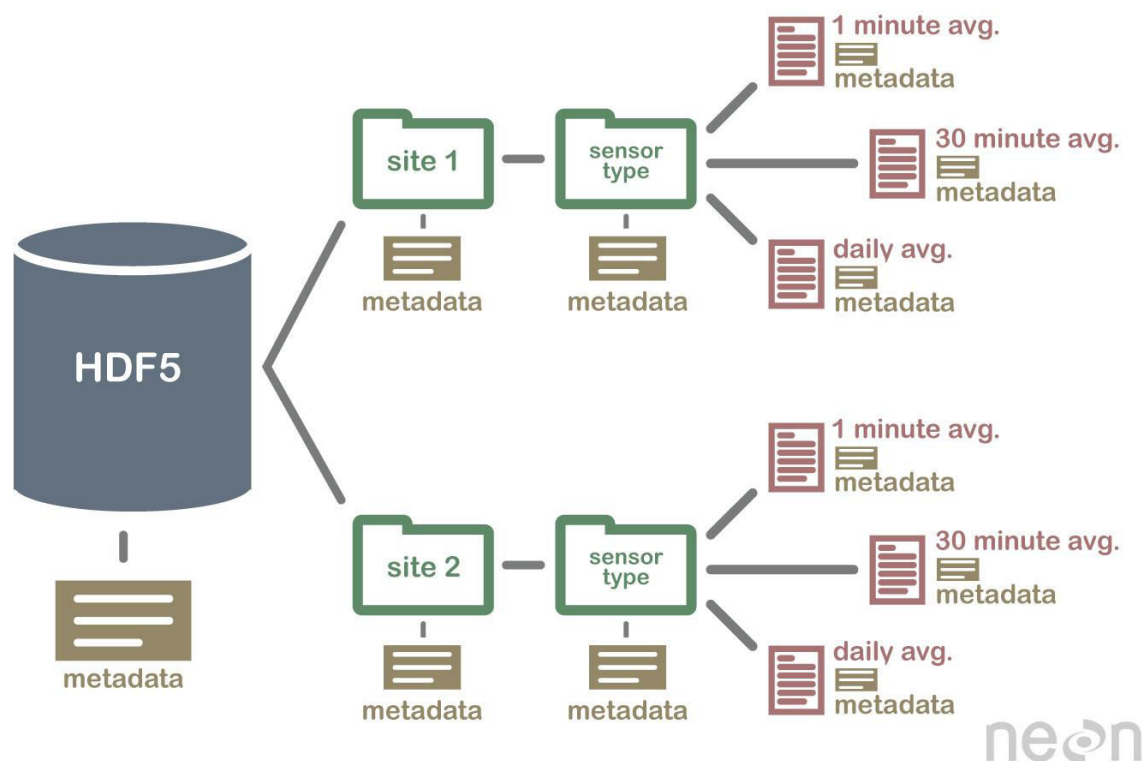
Grupa: Element poput fascikle unutar HDF5 datoteke koji može sadržati druge grupe ili skupove podataka u sebi.

Skup podataka: Su stvarni podaci sadržani u HDF5 datoteci. Skupovi podataka se često (ali nije obavezno) čuvaju unutar grupa u datoteci.



Slika 1 - Primer strukture HDF5 datoteke koja sadrži grupe, skupove podataka i povezane metapodatke.

HDF5 datoteka koja sadrži skupove podataka može biti ovako strukturirana:



Slika 2 - Primer strukture HDF5 datoteke koja sadrži podatke za više lokacija na terenu i takođe sadrži različite skupove podataka (usrednjene u različitim vremenskim intervalima).

HDF5 je samoopisujući format

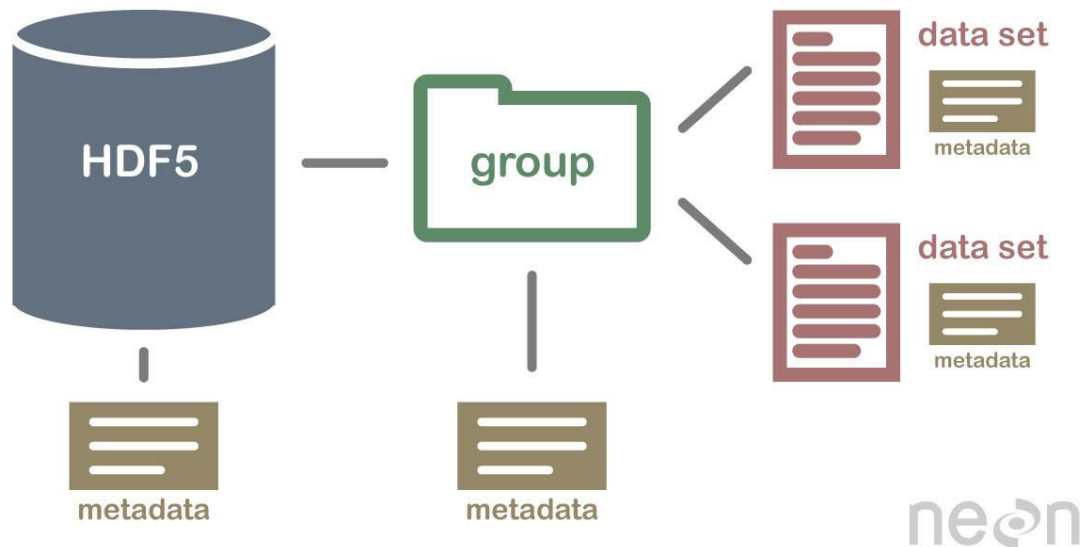
HDF5 format se sam opisuje. To znači da svaka datoteka, grupa i skup podataka mogu imati povezane metapodatke koji opisuju tačno šta su podaci. Prateći gornji primer, možemo da ugradimo informacije o svakom sajtu u datoteku, kao što su:

- * Puno ime i X,Y lokaciju sajta
- * Opis sajta.
- * Bilo koja dokumentacija od interesa.

Slično tome, možemo dodati informacije o tome kako su podaci u skupu podataka prikupljeni, kao što su opisi senzora koji se koristi za prikupljanje podataka o temperaturi. Takođe možemo da priložimo informacije, svakom skupu podataka unutar grupe sajtova, o tome kako je izvršen neki proračun nad podacima i za koji vremenski period su podaci dostupni.

Jedna ključna prednost posedovanja metapodataka koji su priloženi svakoj datoteci, grupi i skupu podataka je da ovo olakšava automatizaciju bez potrebe za posebnim (i dodatnim) dokumentom sa metapodacima. Koristeći programski jezik, kao što je *R* ili *Python*, možemo

da preuzmemo informacije iz metapodataka koji su već povezani sa skupom podataka i koji će nam možda trebati da obradimo skup podataka.



Slika 3 - HDF5 datoteke se same opisuju – to znači da svi elementi (sama datoteka, grupe i skupovi podataka) mogu imati povezane metapodatke koji opisuju informacije sadržane u elementu.

Kompresovana i efikasna podpostavka

HDF5 format je kompresovani format. Veličina svih podataka sadržanih u HDF5 je optimizovana što čini ukupnu veličinu datoteke manjom. Međutim, čak i kada su kompresovane, HDF5 datoteke često sadrže velike podatke i stoga mogu biti prilično velike. Moćan atribut HDF5 je rezanje podataka (data slicing), pomoću kojeg se određeni podskupovi skupa podataka mogu izdvojiti za obradu. To znači da ceo skup podataka ne mora da se čita u memoriju (RAM) i veoma je korisno jer nam omogućava efikasniji rad sa veoma velikim (gigabajtima ili više) skupovima podataka!

Heterogeno skladištenje podataka

HDF5 datoteke mogu skladištiti mnogo različitih tipova podataka unutar iste datoteke. Na primer, jedna grupa može da sadrži skup skupova podataka koji sadrže celobrojne (numeričke) i tekstualne (string) podatke. Ili, jedan skup podataka može da sadrži heterogene tipove podataka (npr. i tekstualne i numeričke podatke u jednom skupu podataka). To znači da HDF5 može da uskladišti bilo šta od sledećeg (i više) u jednoj datoteci:

* Podaci o temperaturi, padavinama i PAR (fotosintetičko aktivno zračenje - photosynthetic active radiation)

* Skup slika koje pokrivaju jednu ili više oblasti (svaka slika može imati specifične prostorne informacije povezane sa njom - sve u istoj datoteci)

- * Višestruki ili hiperspektralni skup prostornih podataka koji sadrži stotine opsega.
- * Podaci sa terena za nekoliko lokacija koje karakterišu insekte, sisare, vegetaciju i klimu.
- * Skup slika koje pokrivaju jednu ili više oblasti (svaka slika može imati jedinstvene prostorne informacije povezane sa njom) i još mnogo toga!

Otvoreni format

HDF5 format je otvoren i besplatan za upotrebu. Biblioteke za podršku (i besplatan pregledač) mogu se preuzeti sa veb stranice [HDF grupe](#). Kao takav, HDF5 je široko podržan u nizu programa, uključujući programske jezike otvorenog koda kao što su *R* i *Python*, i komercijalne programske alate kao što su *Matlab* i *IDL*. Prostorni podaci koji se čuvaju u HDF5 formatu mogu se koristiti u *GIS* i programima za obradu slika uključujući *QGIS*, *ArcGIS* i *ENVI*.

Rezime – prednosti HDF5

Samoopisivanje: Skupovi podataka sa HDF5 datotekom se sami opisuju. Ovo nam omogućava da efikasno izdvajamo metapodatke bez potrebe za dodatnim dokumentom sa metapodacima.

Podrška za heterogene podatke: Različiti tipovi skupova podataka mogu biti sadržani u jednoj HDF5 datoteci.

Podržava velike, složene podatke: HDF5 je komprimovani format koji je dizajniran da podržava velike, heterogene i složene skupove podataka.

Podržava sečenje podataka(Data slicing): „Isecanje podataka“ ili izdvajanje delova skupa podataka po potrebi za analizu znači da velike datoteke ne moraju u potpunosti da se čitaju u memoriju računara ili RAM.

Otvoreni format – široka podrška za mnoge alate: Pošto je HDF5 format otvoren, podržava ga mnoštvo programskih jezika i alata, uključujući jezike otvorenog koda kao što su *R* i *Python* i otvorene *GIS* alate kao što je *QGIS*.

Ekstenzije fajla: .hdf, .h4, .hdf4, .he2, .h5, .hdf5, .he5

Ukratko o formatu Onnx

<https://onnx.ai/>

Šta je format ONNX?

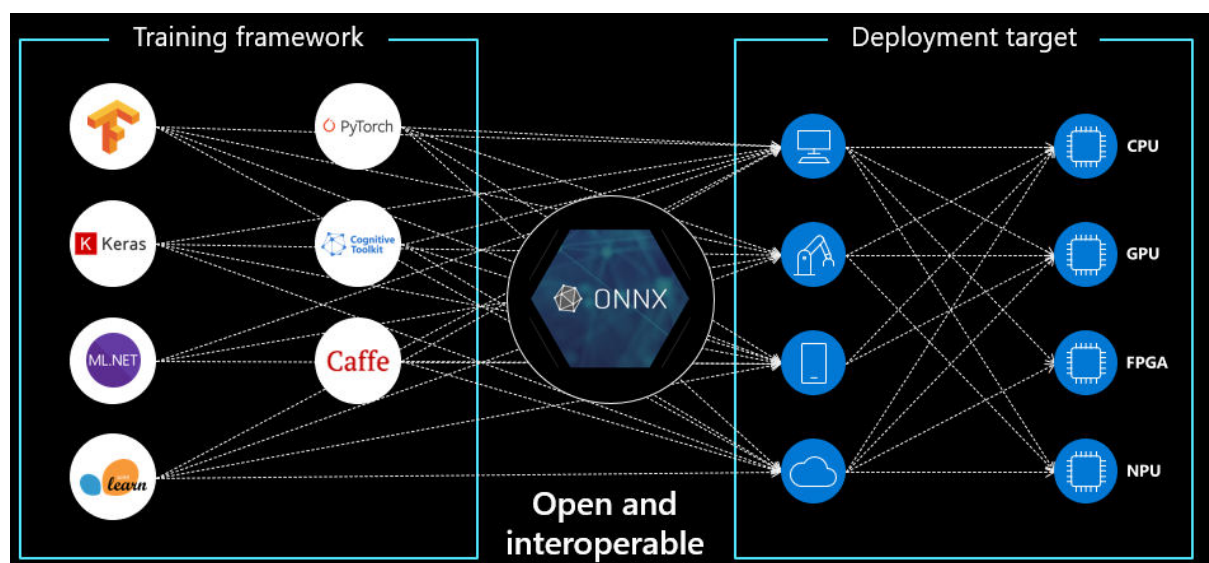
Definicija 1:

ONNX je otvoreni format napravljen da predstavlja modele mašinskog učenja. ONNX definiše zajednički skup operatora – gradivne blokove mašinskog učenja i modela dubokog učenja – i zajednički format datoteke koji omogućava programerima veštačke inteligencije da koriste modele sa različitim frejmvorcima, alatima, vremenima izvođenja i kompajlerima.

Definicija 2

Open Neural Network Exchange (ONNX) je ekosistem veštačke inteligencije otvorenog koda tehnoloških kompanija i istraživačkih organizacija koje uspostavljaju otvorene standarde za predstavljanje algoritama mašinskog učenja i softverskih alata za promovisanje inovacija i saradnje u sektoru veštačke inteligencije.

ONNX je prvobitno nazvan Toffee i razvio ga je PyTorch tim na Fejsbuku. Preimenovan je u ONNX i najavljen od strane Fejsbuka i Majkrosofta.



Različiti frejmvorci dubokog učenja koriste različite vrste grafova. Frejmvorci kao što su *CNTK*, *Caffe2*, *Theano* i *Tensorflow* preferiraju upotrebu statičkih grafova. S druge strane, frejmvorci kao što su *PyTorch* i *Chainer* koriste dinamičke grafove. Ovi grafovi isporučuju srednju reprezentaciju koja hvata tu specifičnu nameru bilo kog izvornog koda. Može da radi na velikom broju uređaja (FPGA, GPU, CPU, itd.).

Pomoćne mogućnosti ONNX-a

Instalacija ONNX-a je isključivo podvrgnuta frejmvorku koji će se koristiti za definisanje modela. Sledeća tabela predstavlja saznanja o mogućnostima podrške ONNX-a -

TYPE OF TOOLS	SUPPORTED TOOLS
Frameworks	Caffe2, Chainer, Cognitive Toolkit, MxNet, PyTorch, Paddle Paddle
Converters	CoreML , MathWorks, Tensorflow
Runtimes	Nvidia, Qualcomm, SOPHON, Tencent, Vespa, Windows
Compilers	Intel AI, skymizer, tvn
Visualizers	NETRON , Visual DL

Koje su prednosti ONNX-a?

Zajednica partnera koji su podržali i razvili ONNX naglašavaju dva kritična faktora kao prednost Open Neural Network Exchange -

- Interoperabilnost okvira – Ovo se može tretirati kao primarna potreba nakon pronalaska ONNX-a. On pruža slobodnu ruku osobama koji se bave podacima da obuče model u frejmvorku i da generišu zaključak u drugom frejmvorku.
- Optimizacije hardvera – Uz korišćenje Open Neural Network Exchange-a, lako je dati mogućnost optimizacije osobama koji se bave podacima. Pruža prednosti ONNX kompatibilnog vremena izvođenja i biblioteka svakom alatu koji koristi izvezeno ONNX modele, i rezultira maksimiziranjem performansi na nekom od najboljih hardvera u tehničkoj oblasti.

Linkovi

Konvertovanje PyThorch modela u ONNX

<https://docs.microsoft.com/en-us/windows/ai/windows-ml/tutorials/pytorch-convert-model>

<https://deci.ai/resources/videos/tutorial-convert-pytorch-onnx/>

Generator koda iz Onnx u PyThorch kod

<https://github.com/fumihwh/onnx-pytorch>

Tensorflow Backend za ONNX

<https://github.com/onnx/onnx-tensorflow>

Uvoz Onnx modela u TensorFlow za izučavanje

<https://thenewstack.io/tutorial-import-an-onnx-model-into-tensorflow-for-inference/>

Konvertovanje TensorFlow modela u Onnx

<https://docs.microsoft.com/en-us/windows/ai/windows-ml/tutorials/tensorflow-convert-model>

<https://onnxruntime.ai/docs/tutorials/tf-get-started.html>

Ukoliko bude bilo potrebe preći sa PyTorch-a u TensorFlow

<https://analyticsindiamag.com/converting-a-model-from-pytorch-to-tensorflow-guide-to-onnx/>