



SALĪDZINOŠĀ ANALĪZE DATU KOPU FORMĀTIEM PYTORCH ATTĒLU KLASIFIKĀCIJAS UZDEVUMIEM

Oļegs Korsaks

Zinātniskais vadītājs:
Dr.sc.ing, Pētnieks
Ēvalds Urtāns

Bakalaura darba priekšizstāvēšana

Tēmas aktuālītāte

- Bakalaura darba tēmas aktuālītāte ir pamatota ar mēģinājumiem paātrināt mašīnmacīšanas procesu. Jo tiek uztaisīts daudz eksperimentu, mēģinot atrast pareizu neironētīkla arhitektūru katram uzdevumam un viss tas aizņem ievērojami daudz laika. Kā arī apmācīšanas procesam jābūt pēc iespējas mērogojamam.

Darba mērķis un tips

- Uz PyTorch piemēra salīdzināt esošos datu formātus, kurus izmanto attēlu klasifikācijas uzdevumos, to ielādes metodes.
- Tips: 1. Moderno risinājumu izpēte

Darba uzdevumi

- Izpētīt MNT pamatus, ConvNet un InceptionNet modeļu arhitektūras, lai apzināti implementēt attēlu klasifikācijas modeļa apmācīšanu.
- Atrast augstas izšķirtspējas datukopas attēlu klasifikācijai.
- Jāizdomā metrikas, kā noteikt ietekmi datu ielādes metodēm.
- Implementēt rīku Python valodā, ar kuru palīdzību dati tiks lejupielādēti, sagatavoti, ielādēti un izmantoti modeļa apmācīšanā. Kā arī izmērīt metrikas.
- Salīdzināt datu kopu formātus un to ielādes metodes dažādam datukopām.

Plānotie rezultāti

- iegūt secinājumus par datu formātiem un to pielietojumu scenāriju salīdzinājumu.

Analītiskā daļa

- Aprakstītas dziļā māšīnmācīšanās pamatarhitektūra, lineārie slāņi, aktivizācijas funkcijas, kļūdas funkcijas, atpakaļizplatīšanas algoritms, apmācamo parametru optimizācijas algoritms, modeļu metrikas, lai dotu saprašanu, cik daudz intensīvu aprēķinu notiek apmācības procesā un kāpēc. Vairāki no tiem tiek pielietoti risinājumā.
- Datu formātu iespēju un īpatnību izpēte ir svarīga, lai būtu iespēja izvēlēties piemērotāko rīku atbilstošam uzdevumam un tā risinājuma veidam.
- Izpētīti HDF5, Zarr, NumPy masīvu npy/npz, JPEG attēlu datu glabāšanai
- Izpētīti PyTorch Dataset un DataLoader, mmap, CuPy, Nvidia DALI + GPUDirect Storage datu ielādes mehānismi

Praktiskā risinājuma daļa

- Implementēts Inception tīkls attēlu klasifikācijai, izmantojot Pytorch
- Implementēta funkcionālitate datu kopu lejupielādei un sagatavošanai (Dataset)
- Izpētīti Nvidia GPUDirect Storage darbības iespējas iegūšanas veidi
- Implementēta funkcionālitate attēlu ielādei
- Iegādāta aparatūra testa videi, testa vides sakonfigurēta

Risinājuma pārbaudes daļa

- Kā praktiskais risinājums ir pārbaudīts?
- ??? Palaižot? :)

Izdarītais

Izdarīts:

- Izpētīt MNT pamatus, ConvNet un InceptionNet modeļu arhitektūras, lai apzināti implementēt attēlu klasifikācijas modeļa apmācīšanu.
- Atrast augstas izšķirtspējas datukopas attēlu klasifikācijai.
- Izvēlēti datu formāti un to ielādes metodes salīdzinājumam.
- Izdomātas metrikas, kā noteikt ietekmi datu ielādes metodēm.

Gandrīz izdarīts:

- Uztaisīta testa vide
- Implementēt rīku Python valodā, ar kuru palīdzību dati tiks lejupielādēti, sagatavoti, ielādēti un izmantoti modeļa apmācīšanā. Kā arī izmērīt metrikas.

Plānots izdarīt pēc gandrīz izdarīta:

- Salīdzināt datu kopu formātus un to ielādes metodes dažādam datukopām.

Vēl plānotais darbs

- Plašāk aprakstīt datu formātus.
- Plašāk aprakstīt datu ielādes metodes.
- Palaist visus testus pēc SSD saņemšanas lai testēšanas apstākļi būtu vienādi.
- Implementēt Nvidia DALI + GPUDirect Storage izmantošanu
- Uzrakstīt rezultātus un tos pamatot.
- Uzrakstīt secinājumus (+rekomendācijas).

Statistika

31 lappuse ir uzrakstīta

Ar vadītāju esmu ticies gandrīz katru otro nedēļu Zoomā visa pavasara semestra laikā

Tik pat reizes esmu sūtījis vadītājam uzrakstītās darba daļas

Paldies par uzmanību!