Title page:
Az adatfeldolgozási témámnak a középpontjában a CAMELS elnevezésű adathalmaz állt, erről fogok ma beszélni valamilyen terjedelemben.

Overview:
- Az előadás tematikája roppant egyszerű
- Először bemutatom a téma hátteréhez szükséges elméleti ismereteket (nagy vonalakban). Tekintve, hogy ez egy asztrofizikai/kozmológiai adathalmaz, így ezen témakörről beszélnék mindenképp először pár szót. (Csak, hogy értsük miről van szó, mi a fizikai motivációja az adathalmaz létrejöttének és a kutatásnak?)
- Következőleg elmondanám, hogy ez a CAMELS adathalmaz pontosan mi is, mi van benne, miért az van benne, ami?
- Aztán beszélnék a technikai részletekről, tehát magáról az adatfeldolgozásról. Hogy csináltam? Hogy valósítottam meg? Miért azt csináltam, amit csináltam? Mik voltak a nehézségek és problémák, amiket meg kellett oldani? (Tehát nyilván amikkel ti is szembenéztetek, csak egészen más témákban.)
- Végül pedig bemutatnám az eredményeimet és hogy milyen konklúziókat lehet ebből levonni

Theoretical background:
- A csillagászatban a kozmológiai témájú kutatások elég szerteágazók és rengeteg témakört ölelnek fel. A legkülönfélébb galaktikus és extragalaktikus megfigyelésektől elkezdve a gravitációs hullám asztrofizikán át az elméleti, valamint numerikus modellezésig bezárólag és még azon is túl, ez egy nagyon aktív és színes kutatási témakör.
- A fő, végső célja viszont hasonló, mint pl. az ugyanennyire színes részecskefizikának, mégpedig a standard modell – egész pontosan itt a kozmológiai standard modell – fejlesztése, pontosítása, kiterjesztése olyan dolgokra, amiket le kéne írnia, de nem teszi.
- Ugye a részfiz. Standard modell célja, hogy leírja a fizikai kölcsönhatásokat, amiket – mint ahogy az a fizika számára ma már jól ismert –, részecskék közvetítenek.
- Ehhez megint hasonlóan egy bármilyen kozmológiai standard modell célja az, hogy leírja (és amennyire lehetséges, meg is magyarázza) a kozmikus nagyságrendeken megfigyelhető jelenségeket és hatásokat. Tehát úgy ezt az univerzumnak nevezett izét nagy vonalakban, hogy kb. mégis mi ez az egész körülöttünk.
- A ma leginkább elfogadott kozmológiai standard modell neve a LCDM modell. Ez a név a Lambda betű és Cold Dark Matter (hideg sötét anyag) rövidítéséből áll össze. Zanzásítva ez a modell annyit állít, hogy az univerzum három összetevőből áll:
    - A sötét energia (ezt jelöljük a Lambda paraméterrel és ez van az LCDM nevében is)
    - A sötét anyag (egész pontosan ennek is a hideg sötét anyag verziója)
    - És a normális, barionos anyag.

Az összes kozmikus jelenség (pl. az univerzum tágulása, CMB, az univerzum nagyskálás szerkezete, tehát a benne levő anyag eloszlása, stb.) pedig ezek GR-vel karöltött hatásainak, valamint kölcsönhatásainak köszönhető.
- Csak még utoljára, hogy újfent párhuzamot vonjak, a részecskefizikai standard modellnek vannak elég súlyos hiányosságai. Pl. az általunk (jelenleg) ismert 4 fizikai kölcsönhatás közül csak 3 szerepel benne, a negyedik teljes egészében ki van hagyva.
- A LCDM-vel pedig szintén elég komoly gondok vannak. Úgy is lehet fogalmazni, hogy ez a modell hasonlóan öregedett jól, mint egy nyári napon felejtett doboz tej. Oké ez lehet picit erős. Azonban azt be kell lássuk, hogy az elmúlt 20-30 évben, a mérési módszereink és a mérőműszereink pontosságának exponenciális fejlődési sebességével egy iramban derült ki, hogy az LCDM a jelenlegi formájában szimplán nem alkalmas a kozmikus jelenségek, elfogadható pontosságú leírására. Tehát nem alkalmas arra, hogy egy valid, kozmológiai standard modell legyen.

- A probléma nyilván, hogy továbbra sincs másik olyan, jobb modell, ami kellően mainstreammé tudott volna válni tudományos körökben, habár azért sorakoznak erre bőven ígéretes jelöltek. És az egész kozmológiával foglalkozó tudományos világ ennek a megoldására van kiéleződve.
- Na most ennek az egész problémakörnek a felderítése és a felmerülő problémák megoldása szintén, nagyon szerteágazó tudományterület. Mindent meg kell vizsgálni, amit lehetséges, legyen az akár elmélet, akár szimuláció, akár mérés, bármi.
- A legtipikusabb problémák a LCDM esetén azok a „parameter tension”-ök.
- A kozmológiai modellek kontextusban minden esetben emlegetett „paraméter” szó az különféle mérhető, származtatható, becsülhető mennyiségeket jelöl (pl. a látható anyag sűrűsége, vagy az anyageloszlás fluktuációinak amplitúdója, stb.). Egy kozmológiai modell pedig ezeknek valamilyen részhalmazát várja bemenetként, amikből egyértelműen tud egy teljes, koherens kozmológiát definiálni. Ezért „paraméterek”, hisz ezek a modellnek a paraméterei nyilvánvalóan.
- Na most a „parameter tension” több dolgot takarhat, de maradjunk az ismert eseteknél. Képzeljünk el, hogy van két fajta méréstípus, amikben mondjuk a Hubble paramétert, a H0-at szeretnénk megmérni. Az egyik méréstípus esetén direktben meg tudjuk ezt mérni, míg a másik esetén megmérünk valamit, majd a LCDM modell segítségével kiszámoljuk a H0-at ezen valami alapján. A probléma, hogy a direktben történt mérés messze hibahatáron kívül nem azt adja eredményül, mint az LCDM-ből kiszámolt érték. Ez egy példa egy tensionre, ami arra utal egyértelműen, hogy a modell nem jó.
- Más, fontosabb paraméterek is szenvednek hasonló problémáktól, sok sebből vérzik ez a LCDM. A CAMELS adatsor segítségével én ebben a projektben az Omega_m (normális anyag sűrűsége) és a sigma_8 (az anyagsűrűség fluktuációinak nagyságát jelölő paraméter) paramétereket vizsgáltam.
- Egyik lehetséges megoldás a kozmológiai paraméterek vizsgálatára – a mérések és elméleti számítások mellett – szimulációk alkalmazása és azok kiértékelése. Ezek azért fontos eszközök a témában, mert szimulálni olyan kozmológiai modellt szimulálunk, amilyet csak akarunk és annyifélét, amennyit csak nem szégyellünk. Vagy legalábbis amennyire tárhelyünk és számítási kapacitásunk van. Emellett szimulációk segítségével – a megfigyelésekkel szemben – különféle kozmológiák statisztikus sokaságát tudjuk megvizsgálni, amik így egy tudományos szempontból roppant értékes, új szemszögből nyújtanak betekintést számunkra a témába.

Description of CAMELS:
- Ez a CAMELS adatsor amivel dolgoztam pedig pontosan ezt, kozmológiai szimulációk egy nagy gyűjteményét tartalmazza. Egész pontosan összesen 4000 valamennyit.
- Ez az adatsor pedig ténylegesen statisztikai / machine learning módszerekkel történő vizsgálatra lett készítve, ahogy a CAMELS betűszó feloldása is mutatja.
- Kozmológiai szimulációkból sokfélét és azokat is sokféleképpen lehet készíteni. Ezekbe most nem is fogok belemenni, csak nagyon nagy vonalakban.
- Kozmológiai szimulációkban mindig valamilyen részecskehalmazt szimulálunk. Ezek vagy tényleges tömegpontokat reprezentálnak és ekkor N-test szimulációkról beszélünk, vagy hidrodinamikai térfogatokat/részecskéket a hidrodinamikai szimulációk esetén.
- A legelterjedtebb módszer ha a szimuláció egy periodikus oldalfalú kockában történik és konvenció szerint ilyenkor valamilyen köbszámnyi darab részecskét helyezünk bele.
- A kiinduló állapot mindig a részecskék valamilyen közel homogén eloszlása, ami a szimuláció során összeomlik és az univerzum nagyskálás szerkezetére hasonlító formát vesz fel. Pl., mint amilyeneket itt is láthattok a képen pár helyen. (Szálas szerkezet.)
- A részecskék konkrét pozícióján és sebességén kívül, a szimuláció típusától függően, egyéb mennyiségeket is ki lehet minden lépésben számolni, pl. a sűrűséget, hőmérséklet, vastartalmat, mágnesezettséget, stb. A részecskék maguk is külön-külön reprezentálhatnak sötét anyagot, csillagokat, gázt, stb. Rengeteg lehetőségünk van.
- A CAMELS adatsor habár 4000 szimulációt tartalmaz összesen, de összesen csak 2,000 olyan van ezek közül, amiben ilyen sok különböző mennyiség is szerepel minden szimulációhoz. Ez a 2,000 db 1,000-1,000 db hidrodinamikai és magnetohidrodinamikai szimulációból tevődik össze.

- A szimulációk 3D-sek, azonban megtehetjük azt a trükköt, hogy feldaraboljuk egy tengely mentén őket szeletekre, majd ezeket a vékony szeleteket összelapítjuk és egy-egy 2D képet készítünk róluk.
- A CAMELS esetén is ez történt. A 2,000 szimuláció mindegyike 15-15 szeletre van vágva és ezekből vannak 2D képek készítve.
- Minden egyes szimulációhoz hozzá van rendelve 6 db paraméter a cél pedig, hogy ezekből a 2D képekből próbáljuk azokat megtanulni valamilyen machine learning módszer segítségével.

Technical details:
- Na én mit csináltam? Tulajdonképpen én az adathalmaz készítőinek egyik cikkét szerettem volna reprodukálni, ami a fent már említett problémát járja körül.
- Ők egy abszolút egyszerű, szokványos formájú, konvolúciós neurális hálót használtak, hogy megbecsüljék az Omega_m és sigma_8 paraméterek értékét, valamint 4 másik paramétert is.
- Ez a 4 másik paraméter bizonyos asztrofizikai effektusokból származó zajokat jelölnek. A cél az lenne, hogy a modell tanulja meg az Omega_m és sigma_8 értékeket és közben szűrje ki teljesen a különféle zajok hatását. Erről szólt ez az említett cikk és ezt próbáltam én is az alapján elérni.
- Ez a kis vizualizáció a cikkben és így általam is alkalmazott neurális háló architektúráját mutatja.
- Ennek a bemenete ezek az említett 2D képek (amik egyébként 256x256 pixel nagyságúak), a modell kimenete pedig egy 12 elemű vektor. Indexek szerint haladva az első hat darab a 2 kozmológiai + a 4 zajt jelölő paraméternek az értéke, míg a következő hat darab pedig az ezekhez tartozó hibák.
- Ahhoz, hogy a modell tényleg ezeket az értékeket tanulja meg, két darab loss függvényt használtak a cikkben. Az első loss függvény a hat becsült paraméter és a valódi paraméterek átlagtól való négyzetes eltérését minimalizálta, míg a második pedig magát a hibát próbálta minimalizálni. A tényleges loss pedig abból állt elő, hogy vettük ezen kettő loss vektor elemeinek logaritmusát, majd a két vektor összegét és végül az értékek átlagát.
- Az egész projektet kétszer csináltam meg, azonban végül Pytorchban született meg a működőképes verzió. Ennek a fő oka, hogy a sokkal inkább elterjedt és kényelmesebb használatú Tensorflow mostanában nagyon fura dolgokat művel, rettenetesen frusztráló a használata és őszintén eléggé elment a kedvem tőle.
- A másik ok, hogy Pytorchban volt a cikkben ismertetett modellhez egy implementáció, szóval azon sokat már nem kellett dolgozni.
- Nagyobb probléma volt, hogy a Pytorch viszont sokkal kényesebb, mint amilyen a Tensorflow (egész pontosan amilyen a Tensorflow VOLT). A Pytorchnak nagyon specifikus formában kell beadni az adatokat és tanítani a benne létrehozott modelleket, amikhez megadott lépéssorozat implementálása szükséges. Emiatt jóval több adatfeldolgozást igényelt, mint sem Tensorflow esetében, ahol annyi az egész, hogy piff adatok skálázása, majd bumm bele a modellbe és kész.

Results:
- Végül pedig az eredményeket mutatnám be. A 2,000 szimuláció egyik felében minden szimuláció-szelet 12, a másik felében levők pedig 13 különböző mennyiséget megjelenítő képet tartalmaztak. A különbség annyi volt, hogy az MHD szimulációk tartalmazták a mágneses tér adatait is.
- Én ezek mindegyikére futtattam 1-1 tanítást és erre az ábrára tettem fel egy példaeredményt. Ez egy jobb eredmény, amit a sima hidrodinamikai szimulációk sötét anyag sűrűséget ábrázoló képei alapján készítettem.
- De itt van pl. egy rosszabb eredmény, ez a magnézium és vas arányát mutató képek alapján készült, szintén a tisztán hidrodinamikai szimulációk felhasználásával.