2. félévi beszámoló

Bagoly Attila (b.attila.93@gmail.com)

Témavezető: Vattay Gábor

2018. 06. 20.

Statisztikus Fizika, Biológiai Fizika és Kvantumrendszerek Fizikája PhD program

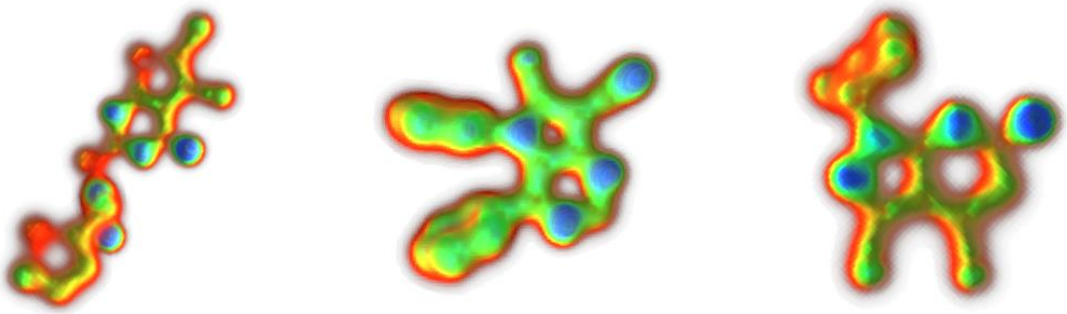
Bevezetés

A félév során az előző beszámolóban bevezetett feladatokkal foglalkoztam, jelentős előrelépéseket érve el. Ezeken felül oktatási tevékenységet végeztem a félév során.

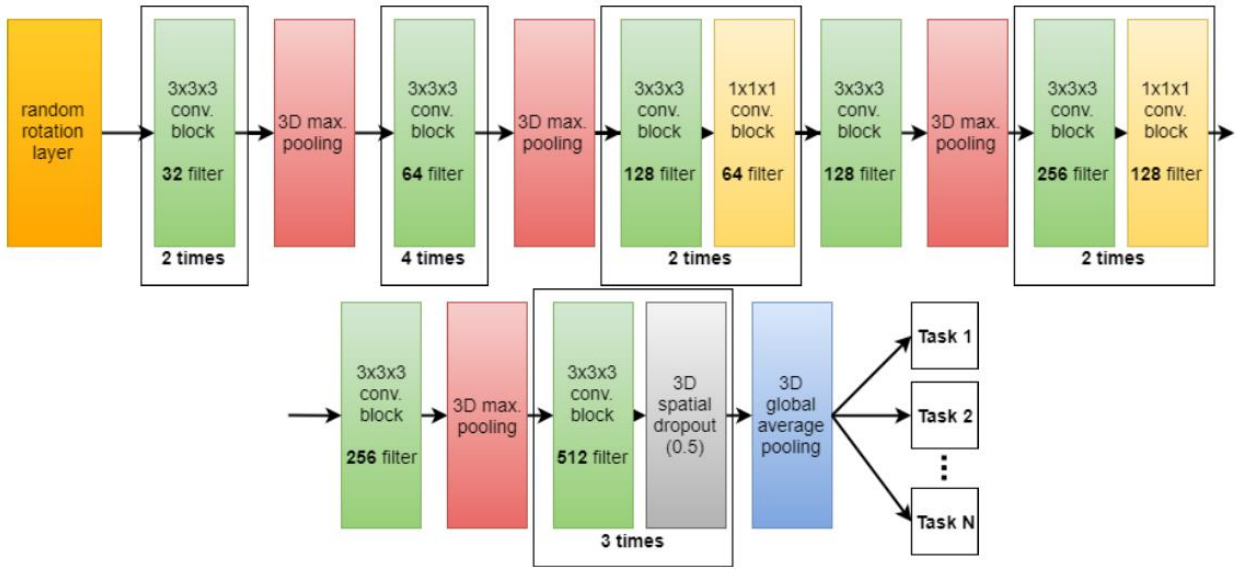
Az aktuális félévben elvégzett kutatások ismertetése

Előző félév végére összeraktam egy egyszerű neuronhálót, amelyet 63000 molekulán előtanítottam arra a feladatra, hogy prediktálja a gyűrűk számát. A bemeneti adat a molekulák töltéseloszlása, amely egy 3D képként fogható fel, amelyre egy 3D konvolúciós hálót építettem. Előző beszámolómban részleteztem az ehhez szükséges lépéseket (TensorFlow operáció CUDA-ban 3D elektronsűrűség számolására, forgatásra stb.), valamint bevezettem a TOX21 adatszetet (10000 molekula, amelynek részhalmazainak aktivitása meg vannak mérve 12 receptorra, amennyiben egy molekula aktivál receptorokat úgy toxikus hatású lesz az ember számára).

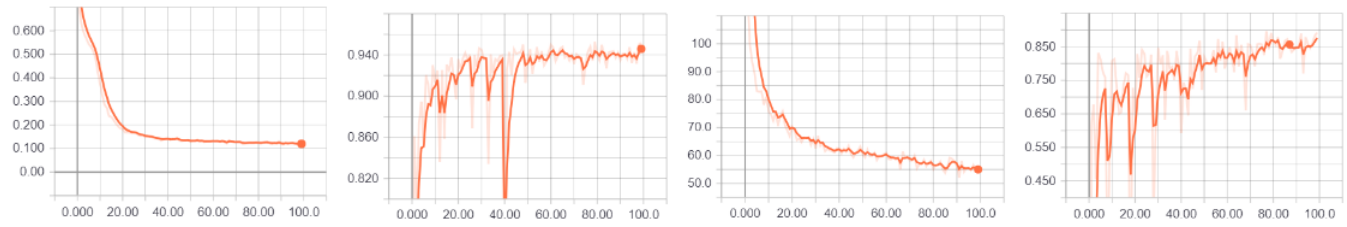
A félév során a molekulák reprezentációja nem változott, továbbra is a 3D elektronsűrűséggel írom le, és erre építek modellt. Néhány molekula esetén ezek a képek 2D projekciója az alábbi ábrán látható:



A félév során egy jóval modernebb neuronhálót fejlesztettem, amely egy teljesen konvolúciós háló. Az optimális architektúra, valamint tanítási paraméterek meghatározása is megtörtént a félévben, ami egy fáradtságos és hosszadalmas munka volt. A legjobb eredményt elérő architektúra az alábbi ábrán látható:



Az új architektúrán kívül az előtanítási feladatot is bonyolítottam, amellyel kapcsolatban az volt a várakozás, hogy majd jobban fog teljesíteni a TOX21 adatszetten. Az alapfelállásban a neuronháló a gyűrűk számának prediktálásával van előtanítva, ami egy aránylag egyszerűbb feladat és könnyen is ellenőrizhető. Ez az egyszerű feladat a sokkal komplexebb aktivitás predikciónál akkor lehet jelentős, amikor felhasználhatók a gyűrűszámoláshoz kialakuló detektorok, azaz az aktivitás valahogy a gyűrűkkel van kapcsolatban. Egy továbbfejlesztési lehetőség más molekula deszkriptorokat is felhasználni az előtanítás során. Ebbe az irányba haladtam a félév során, és a gyűrűk mellett kiválasztottam az úgynevezett topological surface area (TPSA) leírót, amely a molekulák polaritásával kapcsolatos (egy valós szám). A gyűrűkön betanított neuronhálót vettem alapul, amely súlyai a gyűrűkön tanítva lett inicializálva, majd ezt a hálót tanítottam a TPSA prediktálására, miközben továbbra is elvártam, hogy a gyűrűket is jól prédikálja (kényszer). Ez gyakorlatilag azt jelenti, hogy a neuronháló loss függvénye a gyűrűkategória loss függvénye és a TPSA predikció-hoz tartozó loss függvény súlyozott összege. Az alább a két baloldali ábrán látható a gyűrűkhöz tartozó train-loss valamint validation accuracy, a jobboldali két ábrán látható a TPSA loss és az R^2 score a validation adatokon:



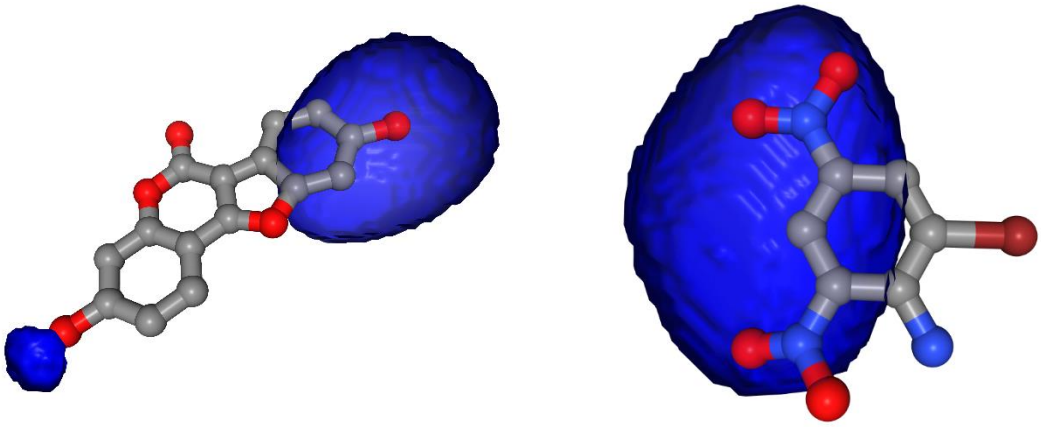
Az így előtanított háló súlyait használva inicializálásként, tanítottam a hálót a TOX21 molekuláinak különböző receptorrokra vonatkozó aktivitás prediktálására. Egy fontos trükk volt a tanítás során, a negatív minták alul-mintavételezése (subsampling), amely során megkerestem az összes olyan molekulát, amely mindenik receptorra inaktív (kb. 2000) és ezek felét véletlenszerűen kihagytam a tanulás során. Az alábbi táblázat mutatja az eredményeket (AUC értékek):

Task	TOX21 winners	FCN	FCN sub-sampling	FCN part-training subsam-pling	no pre-training subsam-pling	pre-train bigger dataset
NR-AR	0.828	0.871	0.864	0.906	0.856	-
NR-AR-LBD	0.879	0.870	0.919	0.876	0.868	-
NR-AhR	0.928	0.793	0.884	0.743	0.880	-
NR-Aromatase	0.838	0.887	0.887	0.790	0.882	-
NR-ER	0.810	0.787	0.825	0.745	0.776	-
NR-ER-LBD	0.827	0.740	0.842	0.695	0.823	-
NR-PPAR-gamma	0.861	0.514	0.729	0.765	0.845	-
SR-ARE	0.840	0.743	0.783	0.681	0.820	-
SR-ATAD5	0.828	0.695	0.805	0.731	0.794	-
SR-HSE	0.865	0.780	0.851	0.613	0.797	-
SR-MMP	0.950	0.838	0.873	0.719	0.874	-
SR-p53	0.880	0.788	0.800	0.694	0.811	-

A táblázat első oszlopa a különböző receptorokat listázza, a második oszlop a hozzátartozó legjobb eddigi eredményeket. A harmadik oszlop az előbb említett neuronháló tanítása, a következő oszlop ugyanezen háló tanítása negatív minták alul-mintavételezésével, a következő oszlop ugyanezen hálót úgy tanítva, hogy alsó rétegéi be vannak fagyasztva (nincsennek illesztve). Az utolsó oszlop pedig mutatja, hogy a háló súlyait random inicializálva milyen eredményeket ér el.

Az eredmények azt mutatják, hogy jelentős a javulás az előző félév végi eredményekhez képest, a mostani eredmények vetekednek az eddigi legjobb TOX21 eredményekkel (ezekhez 40000 molekula deszkriptor volt felhasználva, és sok gépi tanulós módszer ötvözete).

Foglalkoztam továbbá a neuronhálók „értelmezésével” is. Erre az egyik lehetőség, hogy vizsgáljuk, a különböző receptorokra aktívnak mondott molekulák esetén a döntés meghozatalához a neuronháló a molekula mely részeit figyelte. Például az NR-AhR receptor esetén két ilyen ábra az alábbi:



Egy másik vizualizáció, megkeresni, egy adott neuront mely kép-részletek aktiválják legjobban. Ezen vizsgálat során találtam, olyan neuront amely a benzolokra érzékeny és olyat is ami a kálciumra. Azaz a tanítás során kialakult olyan detektor ami elektronsűrűség képéből felismeri a különböző atomokat, alstruktúrákat. Ezen detektorok értékei alapján a magasabb rétegekben levő neuronok már meg tudják mondani, hogy egy molekula aktivál vagy sem egy receptort.

Tanulmányi tevékenység az aktuális félévben

A félév során teljesítettem a Gráfok a bioinformatikában című tárgyat, továbbá deep learninghez és informatikához kapcsolódóan végeztem online kurzusokat (Coursera, Udacity platformokon).

Publikációk

Az eddigi eredmények alapján elkezdtem tudományos publikációt írni, amely várhatóan a következő félévben meg fog jelenni.

Oktatási tevékenység az aktuális félévben

A félév során részt vettem a Csabai István által vezetett *Deep learning és gépi tanulás a tudományokban* című tárgy tematikájának kidolgozásában, készítettem tutorial jellegű részletes házi notebookokat, házifeladatokat kiértékelő szoftvert, valamint weboldalt. Továbbá előadásokat is tartottam a félév során.