FEHÉRJE KLASSZIFIKÁCIÓ FUNKCIÓSOSZTÁLYOK ALAPJÁN

Szökrön Dorottya - Önálló projekt II. (2021/22) - Projektbeszámoló

2022. május 12.

1. Korábbi félév összefoglalása

A tárgy keretein belül a szakdolgozatomban elkezdett munka folytatását, bővítését tűztük ki célul. Ennek témája a gépi tanulás alkalmazása volt molekuláris biológiai feladatokra, konkrétan fehérjék klasszifikálása az elsődleges szerkezetük alapján. A projektmunka során a fehérjék vizsgálatához a természetesnyelv-feldolgozásban (NLP) is használt módszerek közül alkalmaztam néhányat, és összehasonlítottam a különböző, szekvenciákat feldolgozó modelleket.

Az első félév során a szakdolgozatomhoz képest új adathalmaz felhasználása, az abban alkalmazott modellek architektúrájának módosítása valamint új modellek konsturálása történt. Három különböző felépítésű hálózattal dolgoztam:

- Kétirányú LSTM modell
- Konvolúciós modell
- Hibrid modell (előzőek összekapcsolása)

Összességében egyik modell sem teljesítette az elvárásokat. Arra a következtetésre jutottam, hogy a nem megfelelő eremédményekhez hozzájárult az adathalmaz minősége is. Ennek ellenére szükségesnek tartottam további modellek kipróbálását ezen az adathalmazon az eredmények javítása érdekében.

2. 1DCONV & LSTM modell

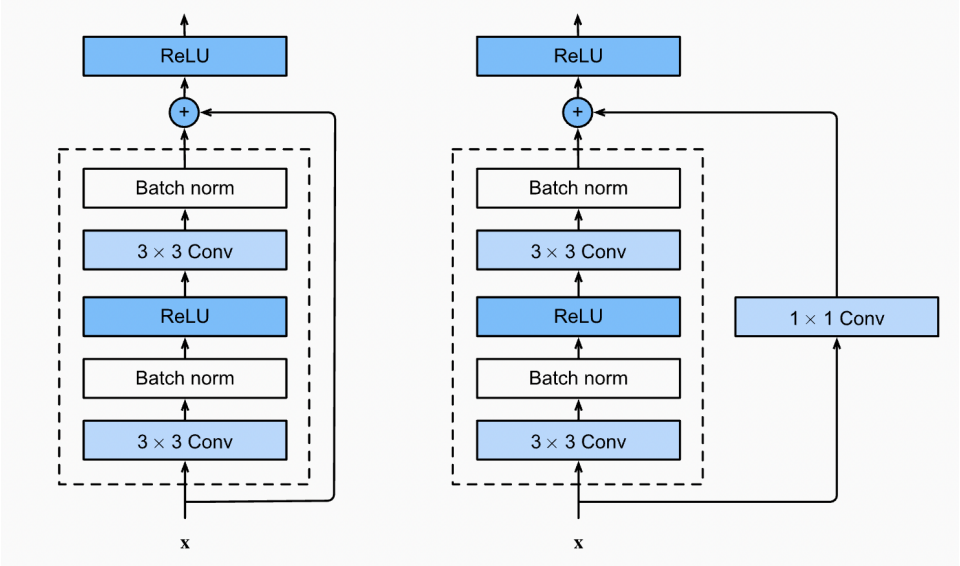
Párhuzamos kutatásban létrejött egy olyan hálózati felépítés, melyet bizonyos fajtájú sztochasztikus folyamatok paraméterbecsléséhez lehetett alkalmazni. A hálózatban az inputként kapott adatok beágyazása egy 1D konvolúciós réteg segítségével valósul meg, majd további feldolgozása kétirányú LSTM rétegek alkalmazásával történik. Ennek mintájára konstruáltam meg a kiinduló modellemet.

3. Residual Networks

A rétegek input adatainak forrását tekintve, az eddigiekben alkalmazott hálózatokról elmondható, hogy minden réteg bemeneti adatai az előző réteg kimeneti adataiból származtak. Ennek az architektúrának a módosításaként reziduális blokkokat alkalmaztam.

A reziduális blokkokat tartalmazó hálózatok abban különböznek a hagyományos felépítésű hálózatoktól, hogy bizonyos rétegek bemeneti adatai nem csak az azt közvetlenül megelőző réteg kimeneti adataiból származnak. Általánosságban elmondható, hogy egy neurális hálózat pontossága javul a rétegek számának növelésével, azonban ez a javulás egy bizonyos határon túl stagnálni, később romlani kezd. Ez a gyakorlatban megfigyelhető jelenség az eltűnő gradiens problémára vezethető vissza.

A ResNet hálózatot 2015-ben mutatták be először, mint egy speciális konvolúciós hálózattípust, ami képfeldolgozásra alkalmazható. A korábbi konvolúciós hálózatoktól eltérően (pl. VGG architektúra) úgynevezett skip connection került a hálózatba. Így valósítható meg az előzőekben említett információáramlás a nem közvetlen egymás után következő rétegek között. Egy LSTM cella belső mechanizmusában is megjelenik a reziduális módszer elve, amely úgynevezett kapuk segítségével valósul meg, ezek szabályozzák az információáramlást. Hasonló megvalósítás lehetséges a rétegek szintjén is.

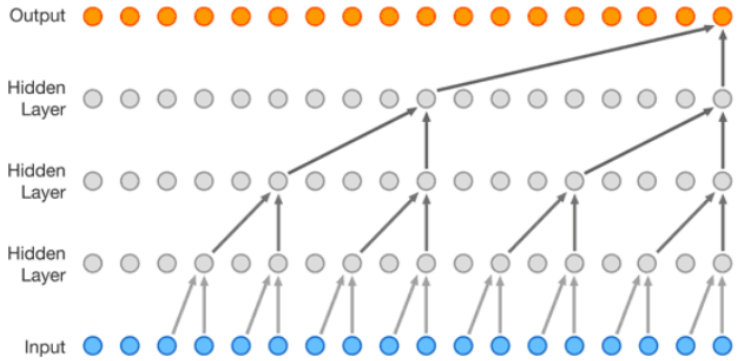


1. ábra. Reziduális módszer

Nincs megkötés a reziduális blokkok belső felépítésére, különböző módokon alkamazható a skip connection is. Például az ábrán látható egyik módszer a kimenet és a reziduális közvetlen összeadása, a másik módszer 1D konvolúció alkalmazása az összeadás előtt a reziduálison, a csatornaszám megváltoztatására.

4. Dilated Networks

A Wavenet hálózat nyers hangadatokat dolgoz fel majd újakat generál. Az egyik erőssége a modellnek, hogy tudja kezelni a nyers hangadatokat, melyek jellemzően 16 000 mintát tartalmaznak másodpercenként, de jobb minőségű adatok esetében ez a szám 22 050 és 44 100 is (22,5/44,1 kHz) lehet. Erre azért képes, mert a modell bizonyos rétegei dilatált konvolúciós rétegek. Hagyományos konvolúció során egy meghatározott méretű konvolúciós szűrővel történik az input vizsgálata, dilatált konvolúciós réteg esetében meghatározható, hogy a konvolúciós szűrő bizonyos pontokat hagyjon figyelmen kívül. Így nagyobb adatrészleteket lehetséges feldolgozni ugyanakkora számítási költséggel. A modell átalakítható klasszifikációs feladat megoldására.



2. ábra. WaveNet struktúra.

A WaveNet modell architekrúrájának kialakítása során az input szekvenciák nagy dimenziója adta az intuíciót. Azonban a dilatációs szűrő azon tulajdonsága, hogy nem csak a közvetlen egymás melletti adatpontokból képes információt nyerni, adta az ötletet, hogy megpróbáljuk ezen módszer előnyeit kihasználni a fehérjevizsgálatok során.

5. Felhasznált modellek

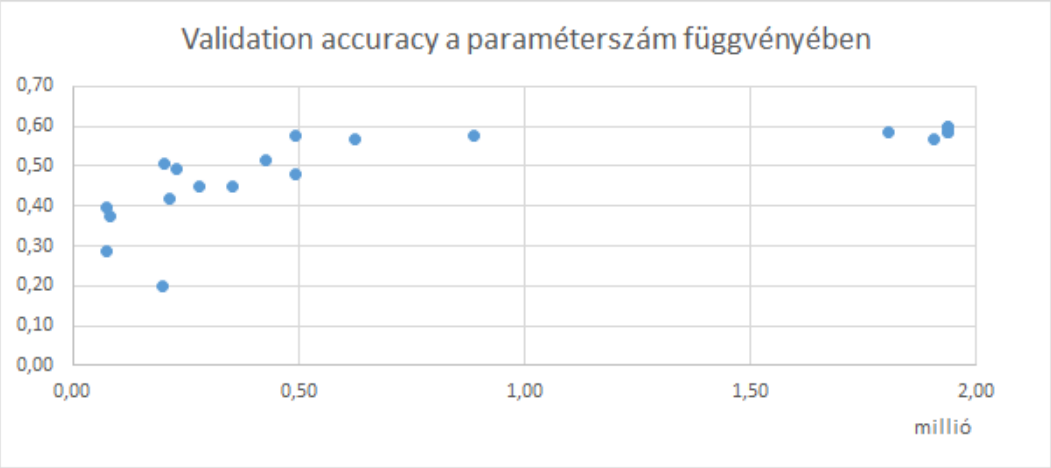
Fontosnak tartom kiemelni, hogy előző félévben végeztem kísérleteket kiegyenlített és kiegyenlítetlen adathalmazon is, azonban úgy határoztam, hogy most csak a kiegyenlített adathalmazzal fogok dolgozni. Különböző forrásokból származó vizsgálati eredményeket összehasonlítva a saját eredményeimmel azt tapasztaltam, hogy strukturálisan szinte megegyező modellek teljesítménybeli különbözőségét csak az adathalmaz befolyásolta. Tehát előfordultak hamisan jó eredményt mutató modellek.

5.1. 1DCONV & LSTM modell

A kiinduló modell a fentebb említett paraméterbecslésre alkalmas modell mintájára jött létre, és sorban egy konvolúciós, egy kétirányú LSTM, valamint egy Average Pooling réteg alkotta. Ezzel a felépítéssel valamint különböző paraméterekkel végeztem a kezdeti méréseket. A modell értékeléséhez az előző félévben alkalmazott modellek legjobb teljesítményét is alapul vettem. A rétegszámok növelése nélkül a modell már az első vizsgálatok során megközelítette a korábbi, már optimalizált modellek teljesítményét, azonos adathalmazon vizsgálva. Valamint jelentős eredménynek tartom, hogy új kétirányú LSTM rétegek

bevezetése után elérte, vagy meghaladta olyan korábbi modellek teljesítményét, amelyek a könnyebb, kiegyenlítetlen adathalmazon dolgoztak. Az LSTM rétegek darabszámának növelése mellett az LSTM cellák neuronszámának növelése által sikerült még javítani az eredményeken. De a szakirodalom és a tapasztalatok alapján elmondható, hogy közvetlen egymás után maximum 3-4 LSTM réteggel tud a modell optimálisan teljesíteni, ennél több réteggel dolgozva a teljesítmény romlani kezd. A probléma megoldása céljából alkalmaztam a reziduális módszert.

5.1.1. Eredmények ábrázolása



5.2. Residual 1DCONV & LSTM modell

A reziduális hálózatok általában konvolúciós rétegeket tartalmaznak, de alkalmazható a módszer LSTM rétegekkel is. Kezdetben a 2. ábrán látható jobb oldali felépítéshez hasonló megoldást alkalmaztam. Konstruáltam egy modellt, mely olyan reziduális blokkot tartalmazott, amelyen belül egy kétirányú LSTM réteg dolgozta fel a blokk bemenetét, és a skip connection a kimenettel való összegzés előtt 1d konvolúción ment keresztül. Ekkor a modell a beágyazó réteg után egy 1D konvolúciós rétegből, majd néhány reziduális blokkból, végül az összegző rétegekből épült fel.

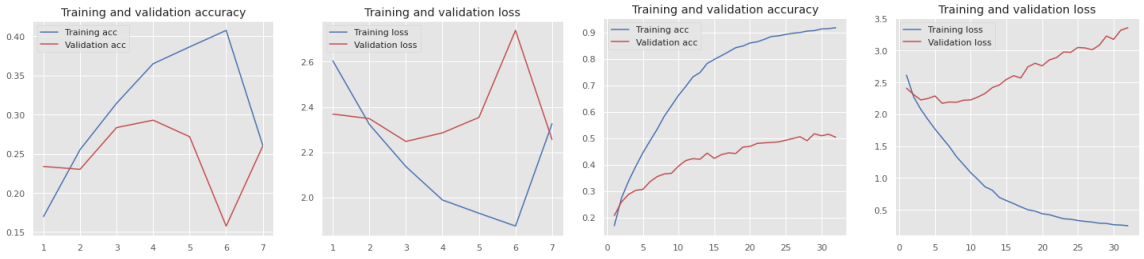
A várakozásaimmal ellentétben ez a modell nagyon rosszul teljesített, ezért úgy döntöttem változtatok a reziduális blokk felépítésén és az összegzés előtti konvolúciót nem alkalmazom. A számítógép erőforrásait figyelembe véve kiválasztottam egy optimálisan jól teljesítő modellt, és annak paraméterbeállításaival dolgoztam. Az LSTM rétegekhez tartozó belső dimenziót 64-nek választottam, és 3-4-5 db LSTM réteggel rendelkező modellek teljesítményét vizsgáltam és hasonlítottam össze.

Összességében elmondható a reziduális blokkot tartalmazó modellekről, hogy a tanulási idő jelentősen kevesebb volt, azonban pontosság szempontjából nem tudtam elérni olyan jelentős javulást, mint amit vártam. De az időtényező nem elhanyagolható különbség, volt olyan modell, amelyiknél közel 4 óráról 1,5 órára csökkent a tanulási idő. A pontosság javítása érdekében ismét módosítottam a reziduális blokkot, és az összegzés művelete helyett konkatenációt alkalmaztam.

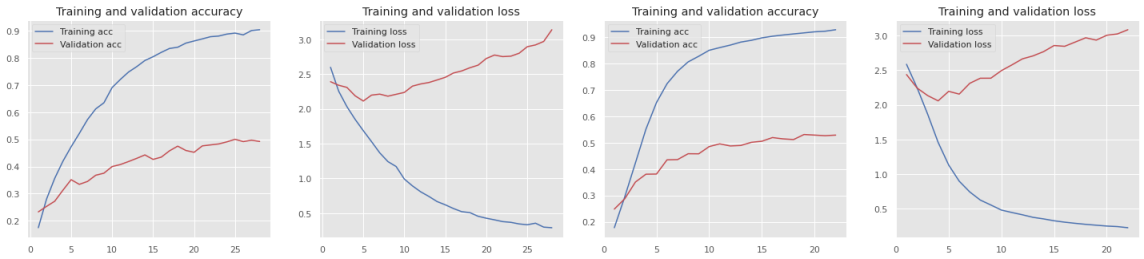
Ismételten az időtényező javulását tapasztaltam, valamint a tanítás során mért validációs veszteség csökkenését. Utolsó módosításként olyan blokkot hoztam létre, amelynek bemenete közvetlenül a beágyazó

rétegből származott és az 1D konvolúciós réteg és a kétirányú LSTM réteg is a blokkon belül volt. Sikerült a pontosságot növelni, azonban a taulási idő jelentősen megnövekedett.

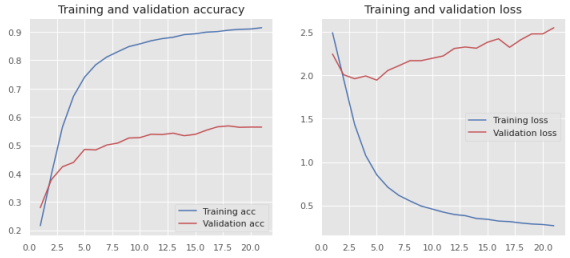
5.2.1. Eredmények ábrázolása



3. ábra. Első típusú reziduális blokk alkalmazása esetén (balra) és az első módosítás után (jobbra)



4. ábra. Reziduális hálózat (balra) és az alapmodell (jobbra), mindkettő 4 LSTM réteggel dolgozott.



5. ábra. Utolsó módosítás utáni, magasabb pontosságú modell.

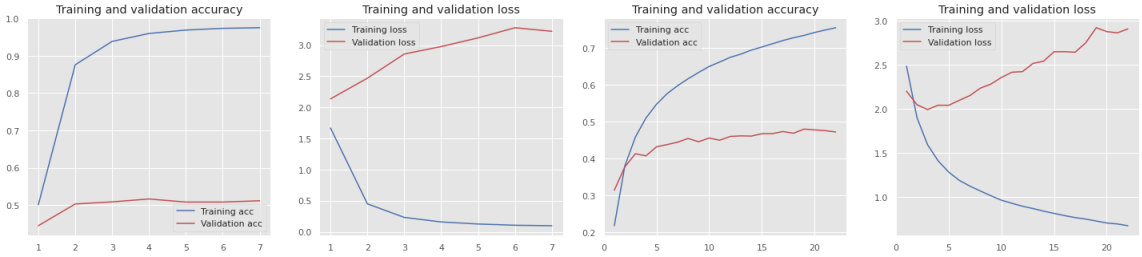
5.3. Dilated 1DCONV & LSTM modell

A fehérjeláncot alkotó aminosavak, nem csak a láncban közvetlenül szomszédos aminosavakkal állnak kölcsönhatásban, ezért arra a következtetésre jutottam, hogy érdemes lehet a távolabbi kapcsolatokat is vizsgálni. A WaveNet modell mintájára átalakítottam az előző félévben használt konvolúciós modellemet dilatált konvolúciós rétegek alkalmazásával. Az elvárásom a modellel szemben a teljesítmény javulása volt a dilatált szűrő miatti bővebb információnyerés hatására.

Az átalakított modellben a beágyazó réteg után egy reziduális blokkon belül valósult meg a dilatált konvolúció. Azt a módszert alkalmaztam, amikor a skip connection 1D konvolúción is keresztül megy. A

dilatációs ráta dinamikus változó lett, mértékét először exponenciálisan növekvőnek választottam, később egyenletesen növekvőre módosítottam az eredmények alapján. A modell pontossága nem javult számottevően, de a loss függvény értéke csökkent.

5.3.1. Eredmények ábrázolása



6. ábra. 1D konvolúciós modell (balra) és WaveNet mintájára létrehozott modell (jobbra)

5.4. Tervek a következő félévre

A következő félévben szeretnék transfomer típusú modellekkel dolgozni ugyanezen a feladaton, valamint új feladatot bevezetni, például fehérjék 3D-s szerkezetének előrejelzését. Szeretnék konstruálni a dilatált konvolúciós réteg mintájára egy dilatált LSTM réteget, amelynek alkalmazása várakozásaim szerint tovább javíthat a modellek teljesítményén.

Hivatkozások

[1] https://www.kaggle.com/shahir/protein-data-set

[2] Dive into deep learning - Zhang, Aston and Lipton, Zachary C and Li, Mu and Smola, Alexander J - 2021

[3] WaveNet: A generative model for raw audio. - Van Den Oord, Aäron and Dieleman, Sander and Zen, Heiga and Simonyan, Karen and Vinyals, Oriol and Graves, Alex and Kalchbrenner, Nal and Senior, Andrew W and Kavukcuoglu, Koray - 2016

[4] How to code your ResNet from scratch in Tensorflow? - This article was published as a part of the Data Science Blogathon - 2021 https://www.analyticsvidhya.com/blog/2021/08/how-to-code-your-resnet-from-scratch-in-tensorflow/

[5] Residual blocks — Building blocks of ResNet - Sabyasachi Sahoo - 2018 https://towardsdatascience.com/residual-blocks-building-blocks-of-resnet-fd90ca15d6ec

[6] Review: DilatedNet — Dilated Convolution (Semantic Segmentation) - Sik-Ho Tsang - 2018 https://towardsdatascience.com/review-dilated-convolution-semantic-segmentation-9d5a5bd768f5