Fehérjék másodlagos szerkezetének prediktálása Acausal Temporal Convolutional Network segítségével

KÉSZÍTETTE: FISCHER KORNÉL

TÉMAVEZETŐ: DR. LUKÁCS ANDRÁS

ÖNÁLLÓ PROJEKT, 2021 DECEMBER

A feladat ismertetése

- Az aminosavak lokális tulajdonsága
- Rengeteg a még annotálatlan fehérje
- A jelenlegi legjobb módszerek homológiára alapulnak, anélkül nem elég hatékonyak
- Olyan modellt keresünk, ami e nélkül is hatékony
- Lewis Moffat és David T. Jones idei cikkéből indultunk ki
- Ők LSTM modellt használtak

Céljaink

- A cikkben szereplő adatbázisok és a cikk eredményeinek reprodukálása
- A háló architektúrájának kicserélése, és ezzel megpróbálni hasonló eredményt elérni

Temporal Convolution Network

- Tulajdonságok
  - Causality
  - Az input bármilyen méretű lehet, és az output ugyanolyan méretű lesz
- Célja, hogy relatív hosszú emlékezete legyen a modellnek

Sequence modeling

- Input: $x_0 , \dots , x_T$
- Output: $y_0 , \dots , y_T$
- Meg kell jósolnunk y_t -t, úgy hogy csak a $x_0 , \dots , x_t$ -t használjuk ehhez
- A sequence modeling network egy olyan f függvény, ami f: $X^{T+1} \rightarrow Y^{T+1}$ és $f(x_0 , \dots , x_T) = y'_0 , \dots , y'_T$
- Szeretnénk megtanulni egy olyan f függvényt, ami minimalizál egy L veszteségfüggvényt:
- $Min_f \; L(y_0 , \dots , y_T, f(x_0 , \dots , x_T))$

Causal Convolutions

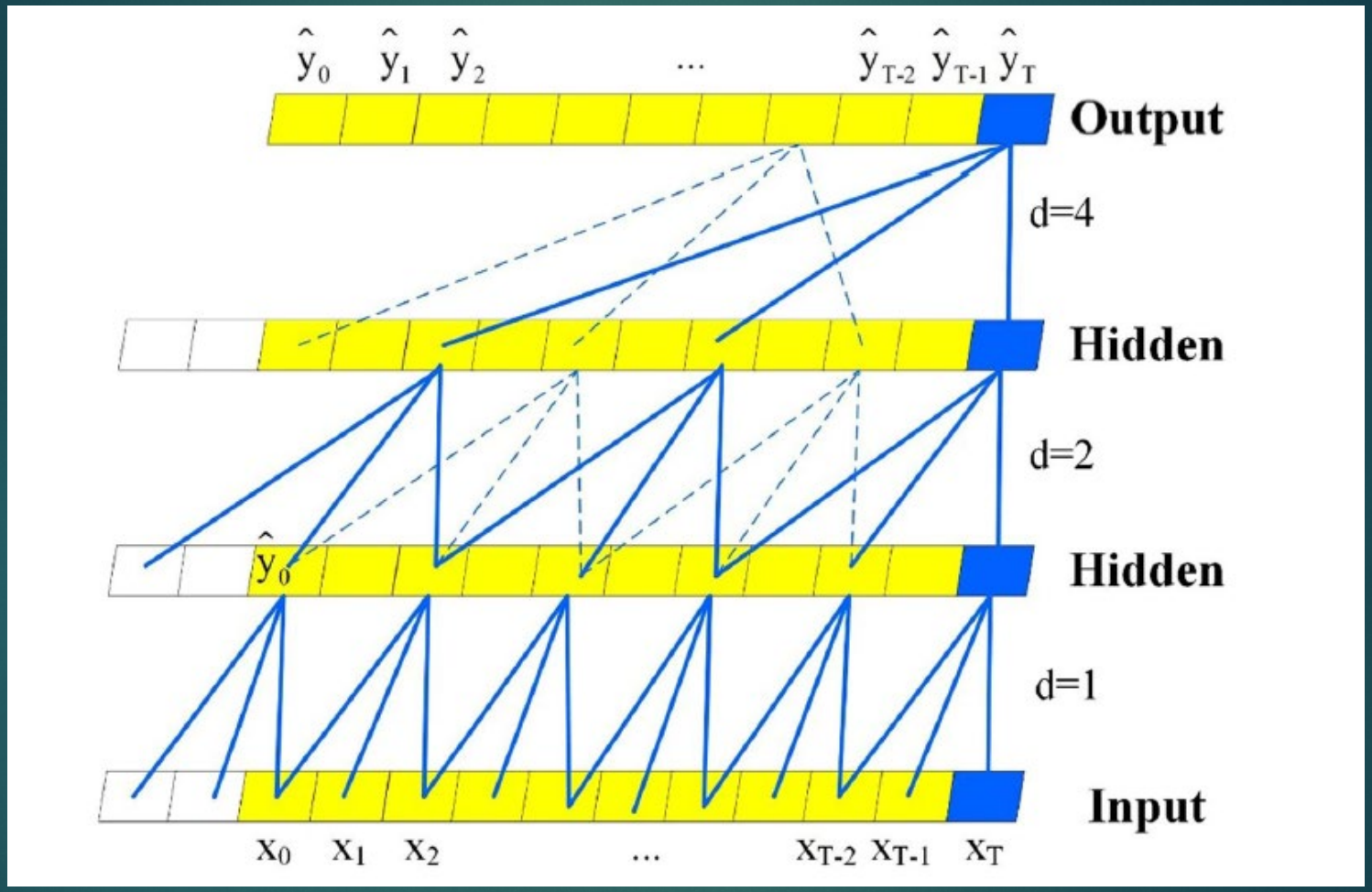
- A causality feltétel teljesítéséhez causal convolution-t használunk
- Azért, hogy az input és az output hossza megegyezzen,1D fully convolution network-öt használunk

Dilated Convolution

- Ha $x \in R^n$ egy szekvencia input, $f: \{0, \dots, k-1\} \rightarrow R$ filter, F egy dilated convolution operátor, s a szekvencia egy eleme

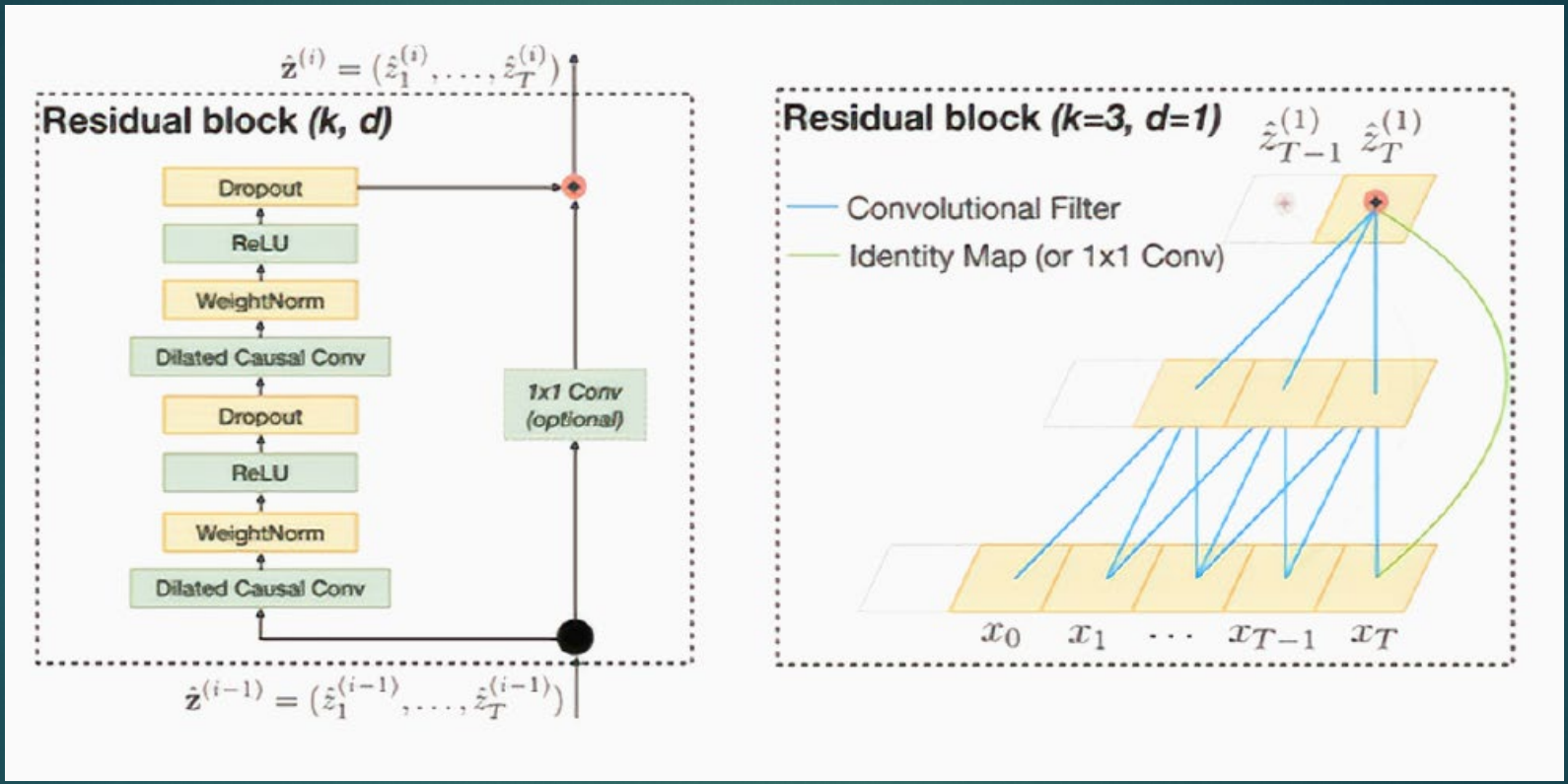
$$F(s) = (x *_d f)(s) = \sum_{i=0}^{k-1} f(i) x_{s-di}$$

Dilation felépítése

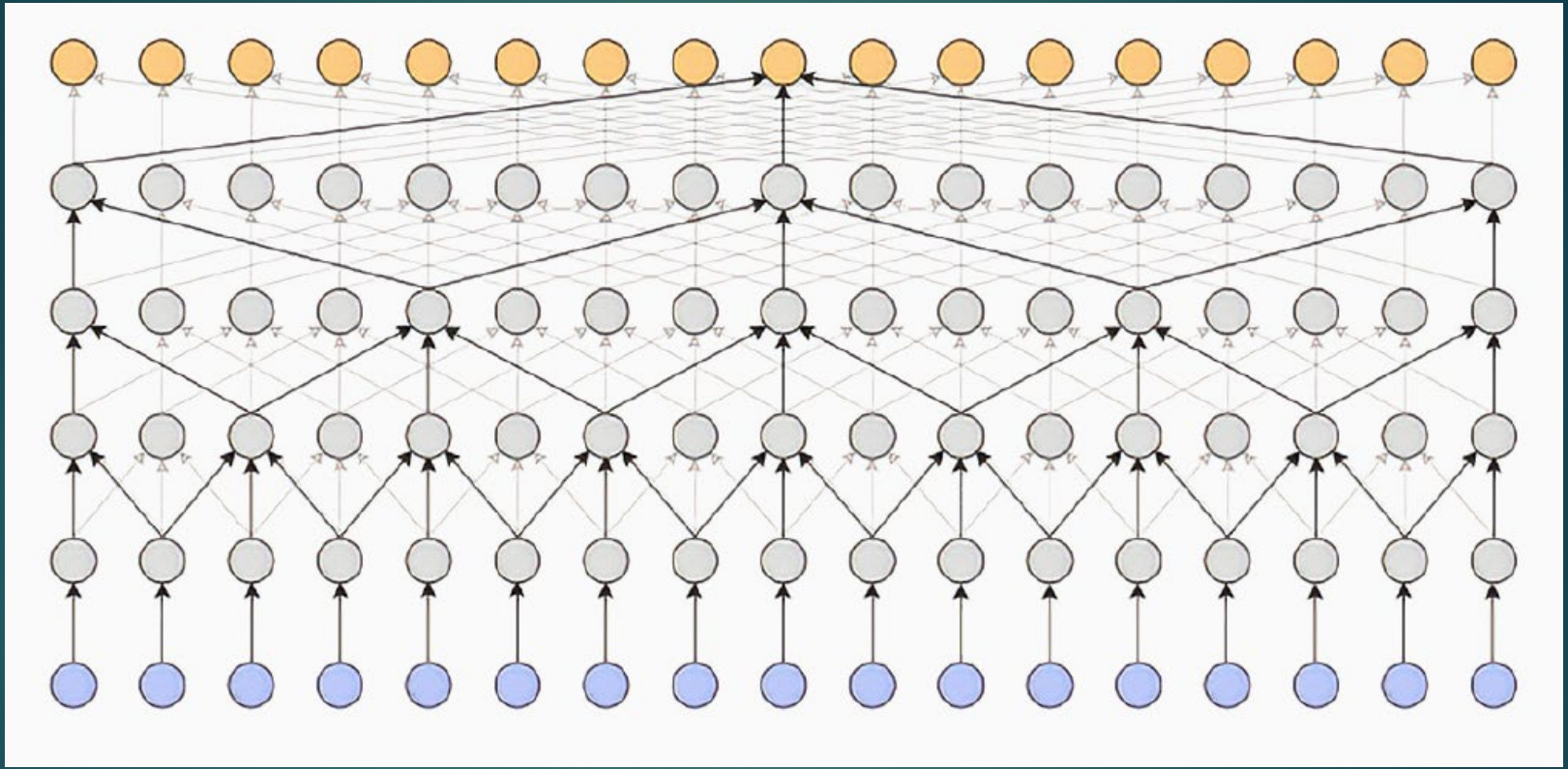


Dilation factors d=1,2,4 és filter méret k=3

Egy reziduális blokk felépítése



Acausal-TCN



Előnyök

- Párhuzamosság
- Könnyen változtatható receptív mező
- Könnyebb gradiens számolás
- Rugalmas inputhossz
- Kevés memória is elég

Az adatbázisok

- UniProt adatbázisból leválogatva 1,08 millió darab szekvencia
- Ezen fehérjékre adtak nagy pontosságú becsléseket a PSIPRED V4 segítségével
- Ezen fehérjéket használtuk a tanításhoz és validációhoz
- Végül a CB513 adathalmazon teszteltük a teljesítményét
- A tanulásnak egy fázisa volt, nem használtunk valódi annotált fehérjéket

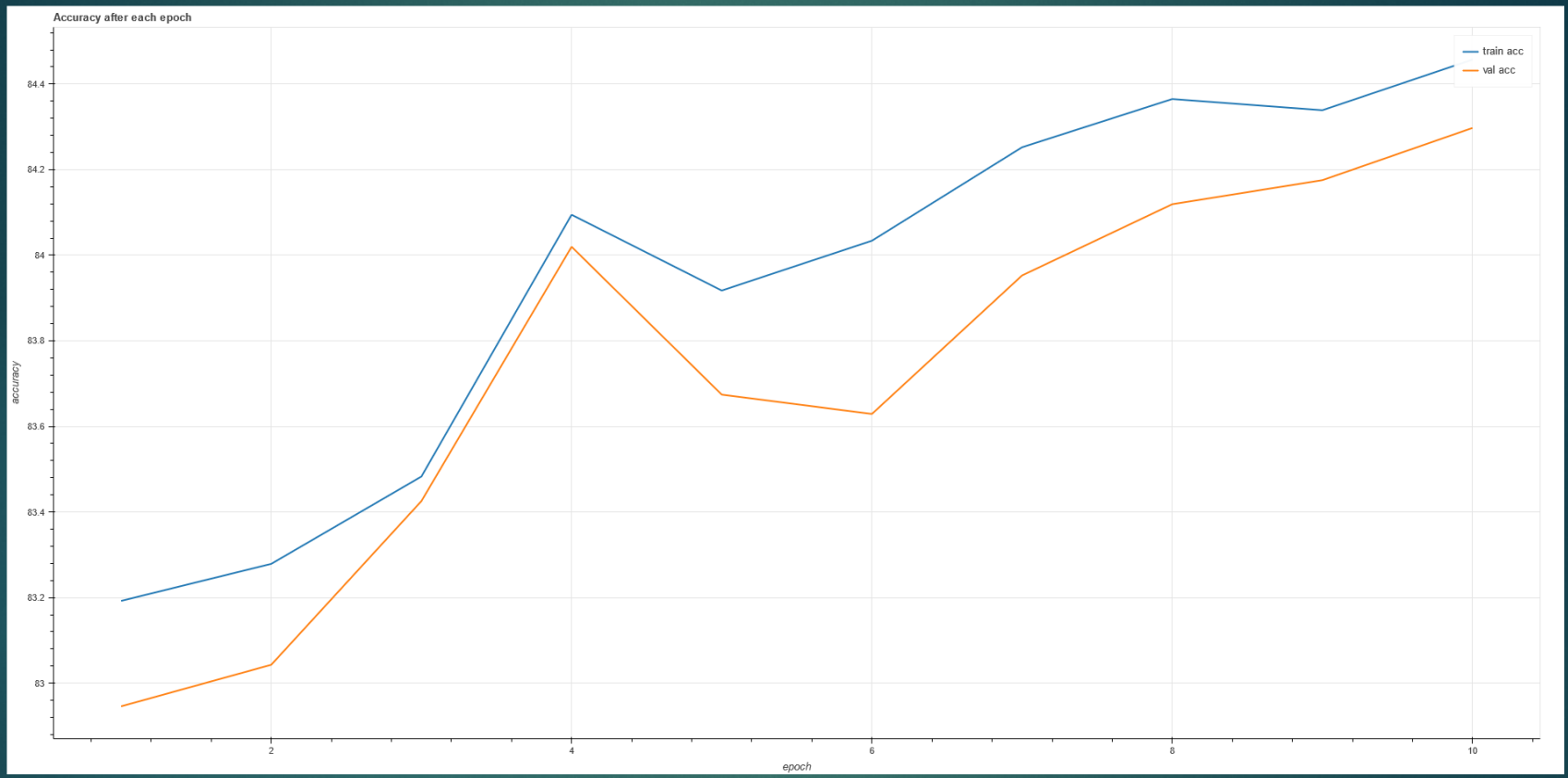
Az adatok előkészítése

- Minden szekvenciát meghagytunk
- Egy szótár segítségével minden aminosavat egy számmal kódoltunk
- Az input vektorok különböző méretűek lettek
- Alkalmaztunk egy beágyazást a háló elején

A háló felépítése

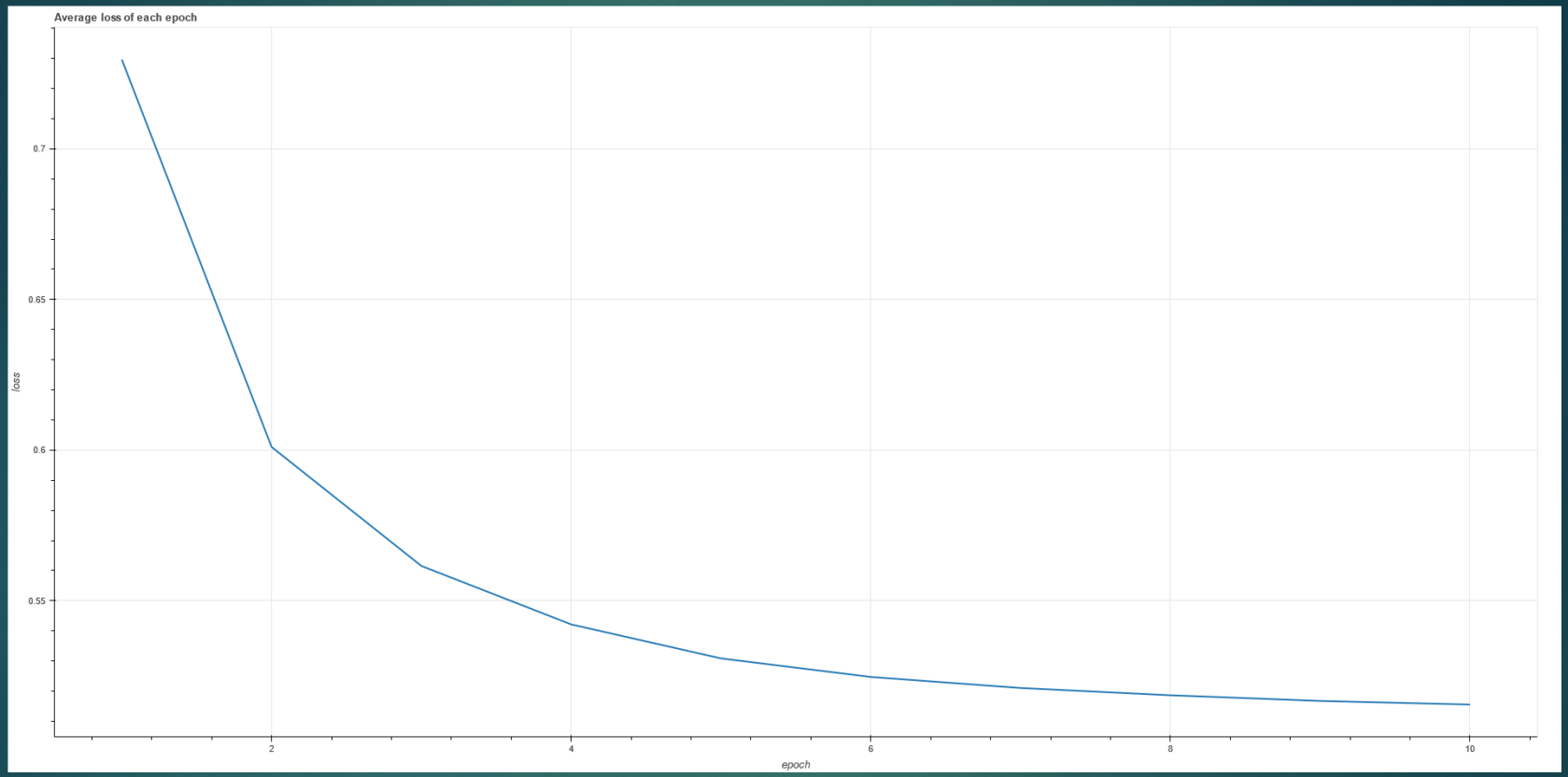
- Egyszerre egy batchméretnyi szekvenciával dolgoztunk
- Az inputot egy tenzorban tároltuk, aminek mérete (batch size)∗(szekvencia-hossz)∗(embedding dimension)
- Adam optimizert használtunk
- A loss függvény kereszt-entrópia volt

Legjobb eredmény



84,45% és 84,29% a pontosság a tanító és a validációs adathalmazon

Legjobb eredmény



A veszteség a 10. epoch végén 0,516 volt

Továbbfejlesztési lehetőségek

- Néhány modell esetén jobb volt a pontosság a validációs adathalmazon, mint a tanítón
- Ennek egy oka lehet, hogy nem volt független a két adathalmaz
- Ezen javíthatunk, ha magunk klaszterezznénk a fehérjéket

Továbbfejlesztési lehetőségek

- A szerzők használták a fine-tuning eljárást a modellük betanítása során, ők javítottak vele
- Mi is megszerezhetnénk az itt használt adatokat, lehet tovább tanulna velük a modell

Köszönöm a figyelmet!