Fehérjék lokációjának meghatározása Long Short Term Memory segítségével

FISCHER KORNÉL

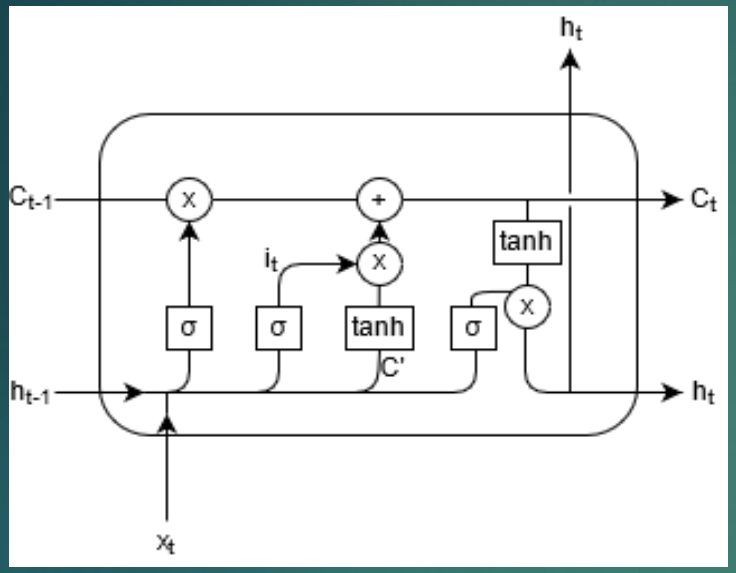
Tavalyi feladat lezárása

- Proteinek kristályosodási hőmérsékletét jósoltuk
- Két modell teljesítményét vetettük össze
- Az XGBoost nem múlta felül a Random Forest teljesítményét
- A tanítások többszöri megismétlésével megnéztük az mse átlagát és szórását
- Ez alapján úgy tűnt, a két modell nagyjából ugyanolyan jól teljesít
- Új, nehezebb feladatot kerestünk

Új feladat ismertetése

- A fehérjék sejten belüli helyének vizsgálata
- Az input az aminosavak sorrendje, az output egy címke
- Rekurrens háló alkalmazása, LSTM

Long Short Term Memory



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

$$o_t = \sigma\left(W_o \; [h_{t-1}, x_t] \; + \; b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

- C: cell state
- x: input
- h: output

Adatok előkészítése

- Előző félévi módszerek tovább javítása
  - Ezúttal meghagytuk a rövid szekvenciákat is
  - Az 1500 aminosavnál hosszabbakat dobtuk ki
- Egy szótár segítségével minden aminosavat egy számmal kódoltunk
- Az input vektorok különböző méretűek lettek
- A rekurrens háló elején alkalmaztunk egy embedding réteget

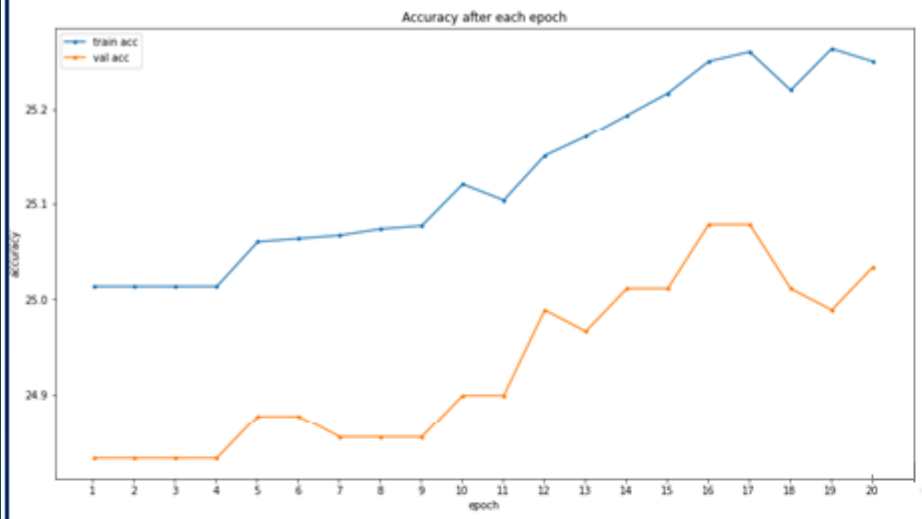
A háló felépítése

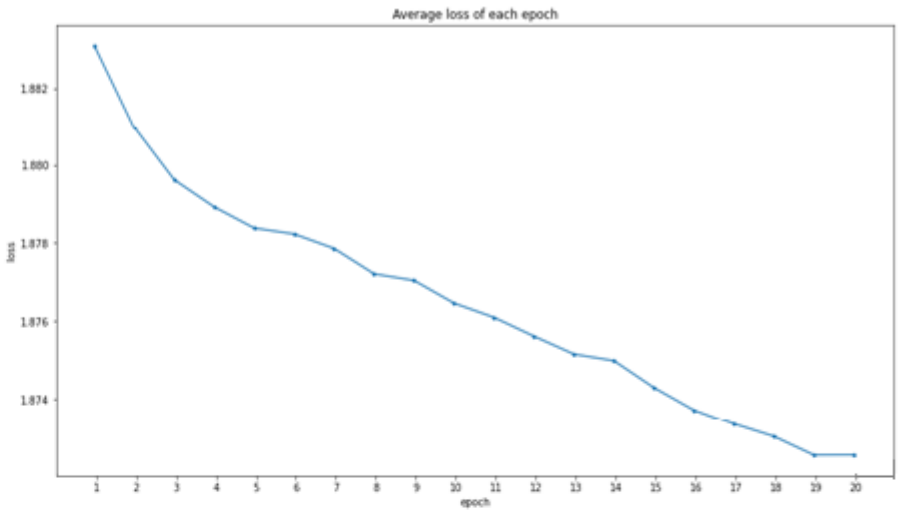
- Egyszerre egy batchméretnyi szekvenciával dolgoztunk
- Az inputot egy tenzorban tároltuk, aminek mérete batchsize∗szekvencia-hossz∗embedding dimension
- Egy fully connected réteggel klasszifikáltunk
- Loss függvénynek kereszt-entrópiát használtunk

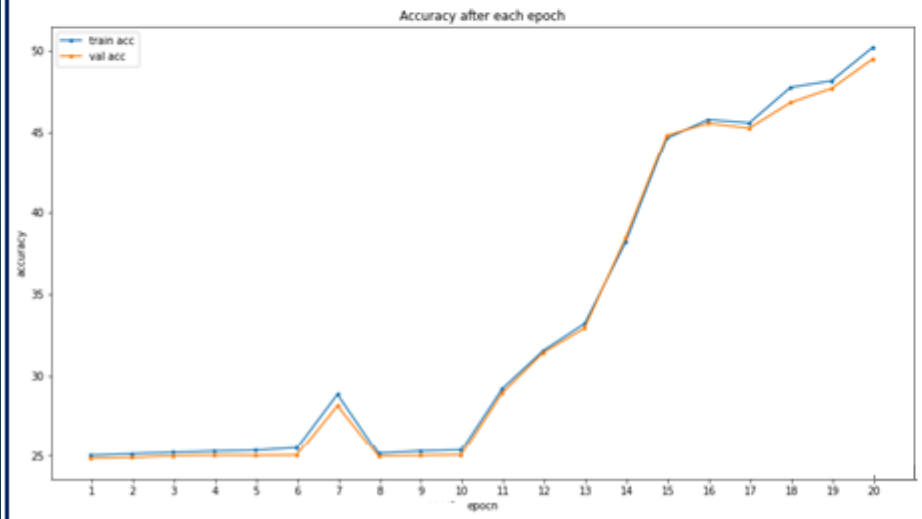
Hiperparaméterek állítása

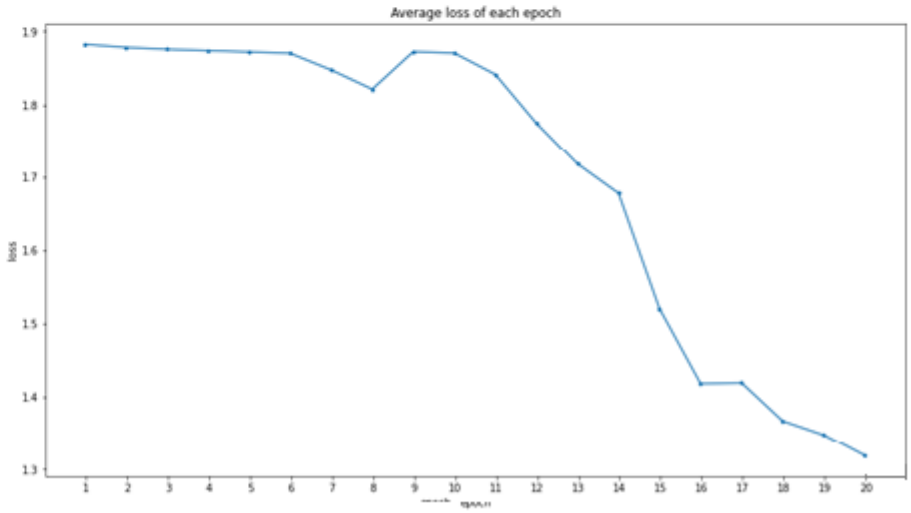
- A lassú futásidő miatt nem volt hatékony a gridsearch
- Kevesebb mérésből kellett jó paraméterezést találni
- Más feladatokból szerzett tapasztalatok szolgáltak iránymutatásul, például a learning rate és a batchsize állításakor

Eredmények kisebb adathalmazon





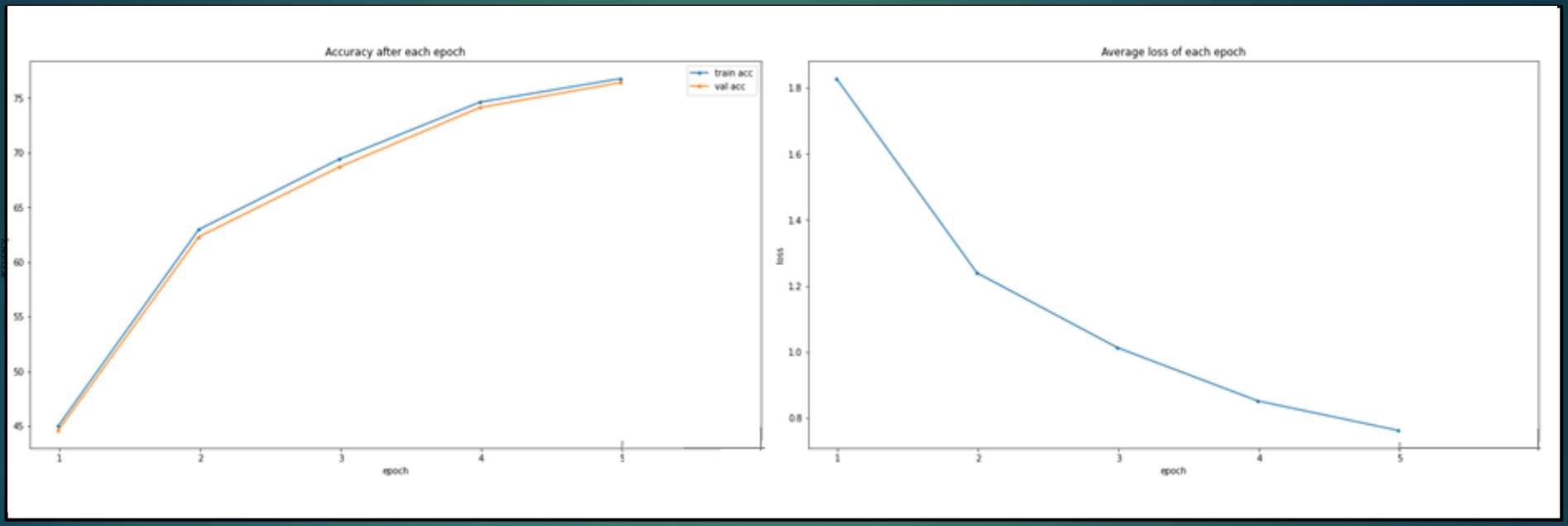




Hiperparaméterek állítása

- Paraméterek és végső értéküк:
  - Learning rate: 0.005
  - Batchsize: 100
  - Embedding dimension: 50
  - Hidden dimension: 100
  - Number of layers: 4
  - Dropout: 0.6
  - Bidirectional: False

Legjobb eredmény



További megfigyelések

- A címkék eloszlása egyenletes volt az adathalmazban
- Az aminosavak sorrendbe tevése nem befolyásolta az eredményt
- Más optimizert is ki lehetett volna próbálni, például SGD-t

Jövőbeli feladatok

- Következő félévben unsupervised tanulás kipróbálása BERT modell segítségével
- Természetes nyelvfeldolgozó technikák alkalmazása a fehérjeszekvenciákra

Köszönöm a figyelmet!