Zadanie 1

Filip Petrán, Jakub Povinec

https://github.com/ns-super-team/assignment-1

Cieľ

Cielom zadania bolo z dostupných dát predikovať, či na základe vstupných atribútov daný pasažier prežil.

Dáta

Dataset je voľne dostupný na:
https://web.stanford.edu/class/archive/cs/cs109/cs109.1176/stuff/titanic.csv.

Dataset, s ktorým sme pracovali obsahuje údaje o pasažieroch lode Titanic, ako napríklad vek pasažiera, trieda v ktorej cestovali (1, 2 alebo 3), pohlavie, meno, počet rodinných príslušníkov na palube alebo cena cestovného. Dataset tiež obsahuje informáciu o tom, či daný pasažier prežil (1) alebo neprežil (0) nehodu.

Vzorka dát

	Survived	Pclass	Name	Sex	Age	Siblings/Spouses Aboard	Parents/Children Aboard	Fare
0	0	3	Mr. Owen Harris Braund	male	22.0	1	0	7.2500
1	1	1	Mrs. John Bradley (Florence Briggs Thayer) Cum...	female	38.0	1	0	71.2833
2	1	3	Miss. Laina Heikkinen	female	26.0	0	0	7.9250
3	1	1	Mrs. Jacques Heath (Lily May Peel) Futrelle	female	35.0	1	0	53.1000
4	0	3	Mr. William Henry Allen	male	35.0	0	0	8.0500
5	0	3	Mr. James Moran	male	27.0	0	0	8.4583
6	0	1	Mr. Timothy J McCarthy	male	54.0	0	0	51.8625
7	0	3	Master. Gosta Leonard Palsson	male	2.0	3	1	21.0750
8	1	3	Mrs. Oscar W (Elisabeth Vilhelmina Berg) Johnson	female	27.0	0	2	11.1333
9	1	2	Mrs. Nicholas (Adele Achem) Nasser	female	14.0	1	0	30.0708

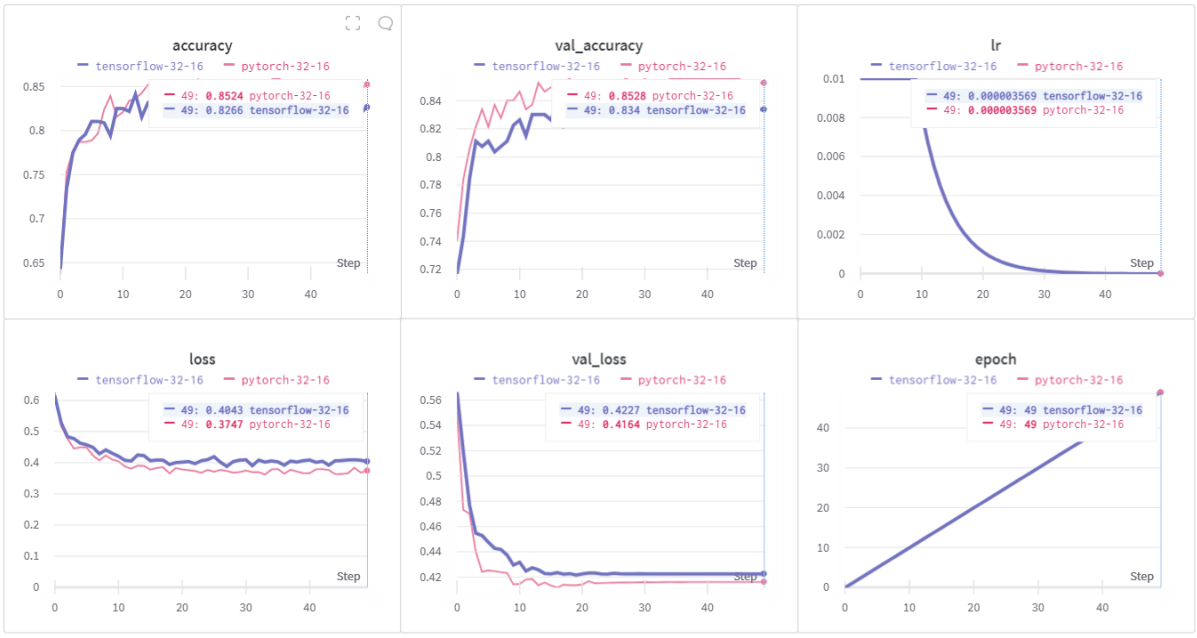
	Survived	Pclass	Sex	Age	Siblings/Spouses Aboard	Parents/Children Aboard	Fare
count	887.000000	887.000000	887.000000	887.000000	887.000000	887.000000	887.00000
mean	0.385569	2.305524	0.354002	29.471443	0.525366	0.383315	32.30542
std	0.487004	0.836662	0.478480	14.121908	1.104669	0.807466	49.78204
min	0.000000	1.000000	0.000000	0.420000	0.000000	0.000000	0.00000
25%	0.000000	2.000000	0.000000	20.250000	0.000000	0.000000	7.92500
50%	0.000000	3.000000	0.000000	28.000000	0.000000	0.000000	14.45420
75%	1.000000	3.000000	1.000000	38.000000	1.000000	0.000000	31.13750
max	1.000000	3.000000	1.000000	80.000000	8.000000	6.000000	512.32920

Zo stĺpcov `Fare` a `Age` sme odstránili outlierov. Stĺpec `Fare` sme taktiež preškálovali pomocou Logaritmického škálovania a na obe stĺpce sme normalizovali pomocou Stadard scalera (odpočítali sme ich priemer a vydelili štandardnou odchýlkou).

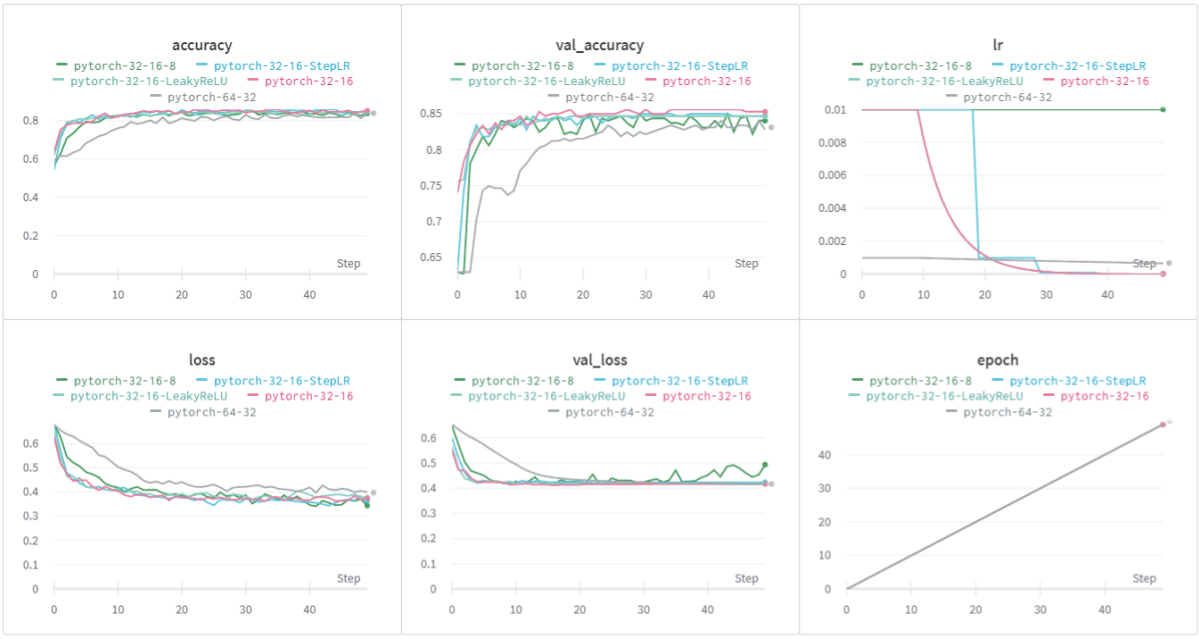
Upravené dáta sme následne rozdelili na trénovací a validačný dataset v pomere 70/30 s tým, že triedy zostali zastúpené v rovnakom pomere.

Model

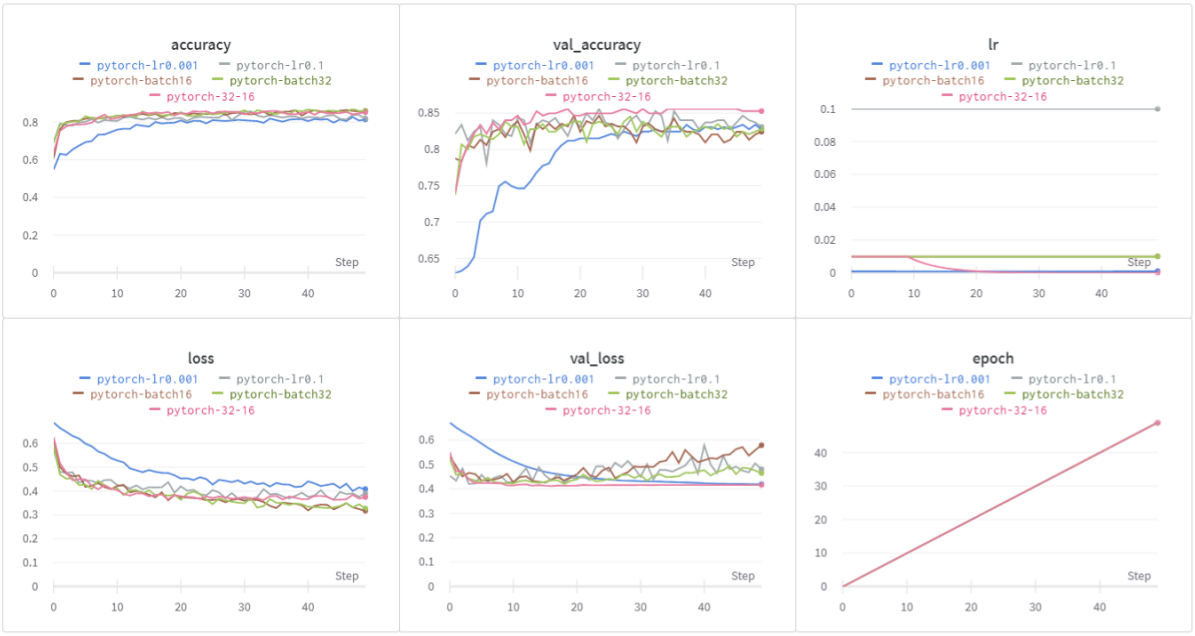
Náš najlepší model mal 2 skryté vrstvy (32 a 16 neurónov), medzi ktorými boli ReLU aktivačné funkcie. Za druhou vrstvou bol taktiež Dropout s pravdepodobnosťou 0.3. Model sme trénovali 50 epoch. Ako optimizer sme použili Adam s 0.01 lr, ktorý sme exponenciálne zmenšovali hodnotou (gamma=0.82).



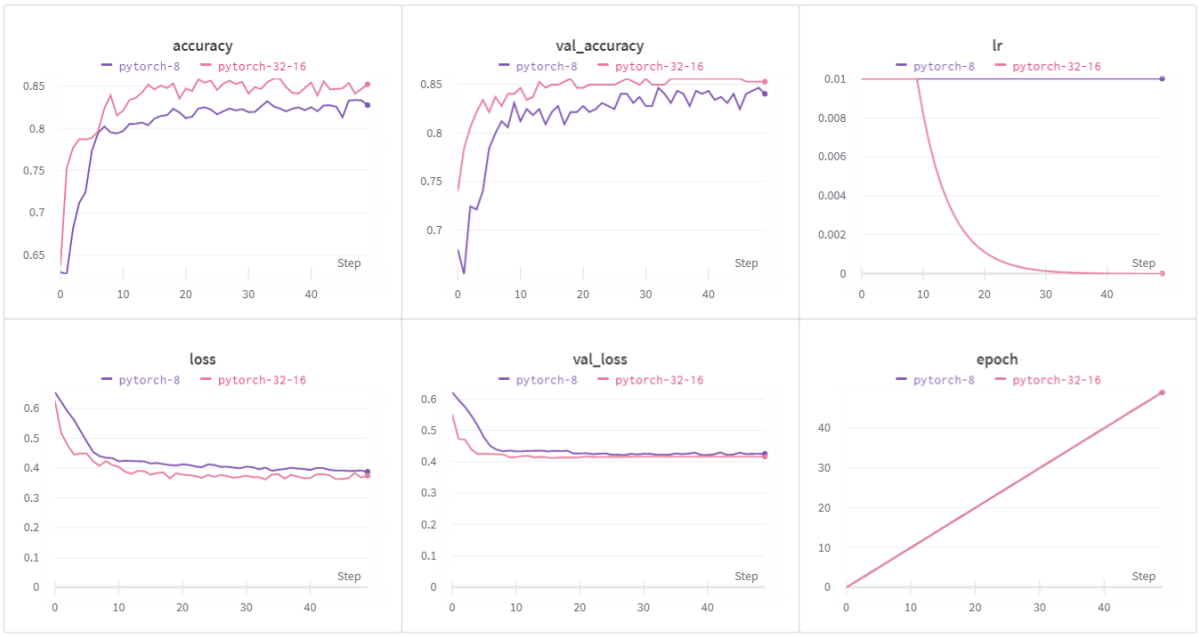
Taktiež sme skúšali použiť iný scheduler (Step), inú aktivačnú funkciu (LeakyReLU) a väčší počet neurónov na jednotlivých vrstvách (64, 32). Pri zväčšení počtu neurónov bolo možné pozorovať väčší overfit, rovnako ako po pridaní ďalšej skrytej vrstvy. Zmena aktivačných funkcií nemala takmer žiadny dopad na výsledky trénovania, podobne ako zmena schedulera.



Vyskúšali sme meniť aj batch size (16, 32) a lr (0.1, 0.001). Zmena batch size mala už výraznejší vplyv na trénovania a modely s menším batchsize boli výrazne overfittnuté.



Najmenší model, ktorý sa nám podarilo urobiť mal 1 skrytú vrstvu s 8 neurónmi a bez dropout-u a LR schedulera, ktorý sme v tomto prípade nepotrebovali, keďže mal model malú kapacitu a ne-overfittoval.



Zhodnotenie

Náš najlepší model dosiahol **validačnú accuracy 0.8528** a **validačný loss 0.4164**.