

Prípadové štúdie dolovania údajov – projekt

Riešiteľ: Bc. Matúš Revický Im1

Zadanie:

Nájsť si dataset aj s úlohou (odporúčaným zdrojom pre všetkých študentov na predmete bol Kaggle), kde by sa dali aplikovať neurónové siete.

Pozn.

- Zadanie bolo dané ústne. Jedná sa o voľnú formuláciu.
- Na začiatku bolo odporúčané použiť fastai2 dostupné na <https://course.fast.ai/>, kde sa nachádzajú tutoriály vo forme videí aj colab notebookov
- Odporúčané použiť Colab

Moje riešenie:

1. Nájdenie vhodného datasetu

Mal som absolvovaný predmet úvod do neurónových sietí. Síce som mal už viacero softvérových projektov, ale žiadne skúsenosti s návrhom neurónových sietí podľa daného problému.

Viac-menej náhodne som sa rozhodol pre dataset AGE, GENDER AND ETHNICITY (FACE DATA) dostupný na <https://www.kaggle.com/nipunarora8/age-gender-and-ethnicity-face-data-csv>

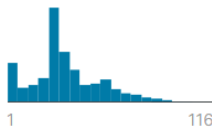


# age	# ethnicity	# gender	▲ img_name	▲ pixels
Age of the person in the image	Specifies the ethnicity of the person	Gender of the person	Name of the Image	Array to String of the image pixels
			23479 unique values	23315 unique values
1	2	0	20161219203650636.jpg g.chip.jpg	129 128 128 126 127 130 133 135 139 142 145 149 147 145 146 147 148 149 149 150 153 153 153 152 153 ...
1	2	0	20161219222752047.jpg g.chip.jpg	164 74 111 168 169 171 175 182 184 188 193 199 200 199 200 196 198 192 193 188 187 186 187 188 183 1...

Figure 1 Ukážka dát

Dataset obsahuje grayscale obrázky tvári o rozmere 40x40 pixelov. Je ich cca 23 000 a obsahujú aj labely.

2. Otestovanie dostupných notebookov

V tomto čase ešte neboli na spôsob implementácie kladené žiadne nároky. V čase testovania bolo na kaggli asi 10 notebookov. Všetky boli v Kerase a v skutočnosti sa jednalo iba o jeden 10 krát skopírovaný notebook, kde bol pre každý model použitý samostatný model (single task). Tento notebook som si skopíroval a otestoval. Prešiel bez errorov. Dĺžka trvania bola relatívne krátka. (asi 5 minút, aj nahrávanie do google drive trvalo krátko (cca 15 min))

3. Odporučené použiť adversial learning prístup

Na Kaggli nebol k dispozícii notebook, kde by bol použitý adversial learning pri podobnom probléme. Niekoľko hodín som o tom čítal články, ale pri ďalšej hodine, sme dospeli k záveru, že je to nad rámec tohto 4 kreditového predmetu z C-bloku.

4. Odporučené použiť fastai2 na návrh vlastného modelu

Jediné materiály, ktoré k fastai2 existovali boli dostupné na <https://course.fast.ai/>.

Citácia z <https://forums.fast.ai/t/fastai2-and-new-course-now-released/75684> :

*“fastai v2 and the new course were released on **August 21st 2020**. fastai v2 is not API-compatible with fastai v1 (it's a from-scratch rewrite). It's much easier to use, more powerful, and better documented than v1, and there's even [a book \(624 pages!\)](#) about it. The book is also available for free as [Jupyter notebooks](#). fastai v2 is documented here: <http://docs.fast.ai/>.”*

Pozn. Nejednalo sa priamo o obrázky, ale o polia stringov, takže bežné knižnice, ktoré výrazne zľahčujú prácu s obrázkami sa nedali použiť. Rovnako sa nedalo jednoducho použiť ani predpripravené modely, ktoré potrebujú obrázky s tromi farebnými kanálmi. Práca s predpripravením dát bola netriviálna.

Po zhliadnutí viacerých tutoriálov a vyskúšaní viacerých notebookov som vytvoril prvý použiteľný model pre rozpoznávanie pohlavia. Po pravidelnom hlásení progresu som ho ešte doplnil na 4 konvolučné vrstvy, upravil dropout... Napísané už bolo aj vyhodnotenie modelu. Následne som podobne napísal aj model na predikciu rasy (multilabel klasifikácia). Riešenie bolo funkčné a napísané vo fastai2. Presnosť pre rasu bola 80% a pre pohlavie 90% pri použití fit one cycle asi už po 10 epochách. Učiaci polomer pre fit one cycle som upravil na hodnotu 0.06.

Príloha:

- **Working_singletask_fastai_v2_custom_implementation.ipynb**

V prílohe je funkčná verzia singletask modelu.

Pozn. Počas práce som došiel k zisteniu, že fastai2 má **veľké problémy so spätnou kompatibilitou** aj medzi minoritnými verziami. Stalo sa mi, že po minoritnom update fastai2 už nebolo možné používať niektoré parametre.

5. Určené použitie multitask modelu a fastai2

Napriek tomu, že teoreticky som vedel presne čo urobiť, vo fastai2 to však hlásilo viacero chýb. Ani po 10 hodinách strávených debugovaním sa chybu nepodarilo odstrániť. (viď. `Testing_multitask_fastai_v2_custom_implementation.ipynb`)

Preto som sa rozhodol otestovať 2 multitask modely vo fastai1 podľa článkov:

- <https://towardsdatascience.com/multi-task-learning-with-pytorch-and-fastai-6d10dc7ce855>
- <https://zhang-yang.medium.com/multi-task-deep-learning-experiment-using-fastai-pytorch-2b5e9d078069>

Presne som zreplikoval postup týchto dvoch článkov s použitím fastai1. Oba končili errorom, podobne ako pri mojej fastai2 implementácii. Teda som dospel k záveru, že fastai1 má podobný problém ako fastai2 a to problémy s **kompabilitou aj medzi minoritnými verziami**.

Prílohy:

- `Testing_multitask_fastai_v1_article1.ipynb`
- `Testing_multitask_fastai_v1_article2.ipynb`
- `Testing_multitask_fastai_v2_custom_implementation.ipynb`

Po otvorení je vidieť, kde výpočet končí errorom.

Pozn. Testovanie zabralo netriviálne množstvo času. Problémom je už samotné nahratie cca 23000 obrázkov. Ak už sú nahraté na drive, tak vyskakuje `drive-timeout error`, takže asi 5 minút pri resetovaní notebooku trvá len načítanie dát. Factory reset notebooku je nutný ak vyskočí `CUDA error`. Colab je v rôznych časoch rôzne zaťažovaný. Niekedy jedna epocha trvá 8 sekúnd, inokedy 800 sekúnd.

6. Multitask Keras

Dataset som nahradil zložitejším <https://www.kaggle.com/jangedoo/utkface-new>. Pre Keras uvádzam notebook, ktorý už obsahuje všetky potrebné informácie.

Projekt rieši 3 problémy:

- **Predikcia pohlavia (`binary_crossentropy`)**
- **Predikcia rasy (`categorical_crossentropy`)**
- **Predikcia veku (`regression`)**

Pozn. Nový dataset obsahuje obrázky o rozmere 200x200 s tromi farebnými kanálmi, učenie je výrazne pomalšie ako pri obrázkoch 40x40 s jedným farebným kanálom vo fastai2. (celý beh bol asi 12 hodín pri 100 epochách)

Pozn. Interaktívne grafy

Prílohy:

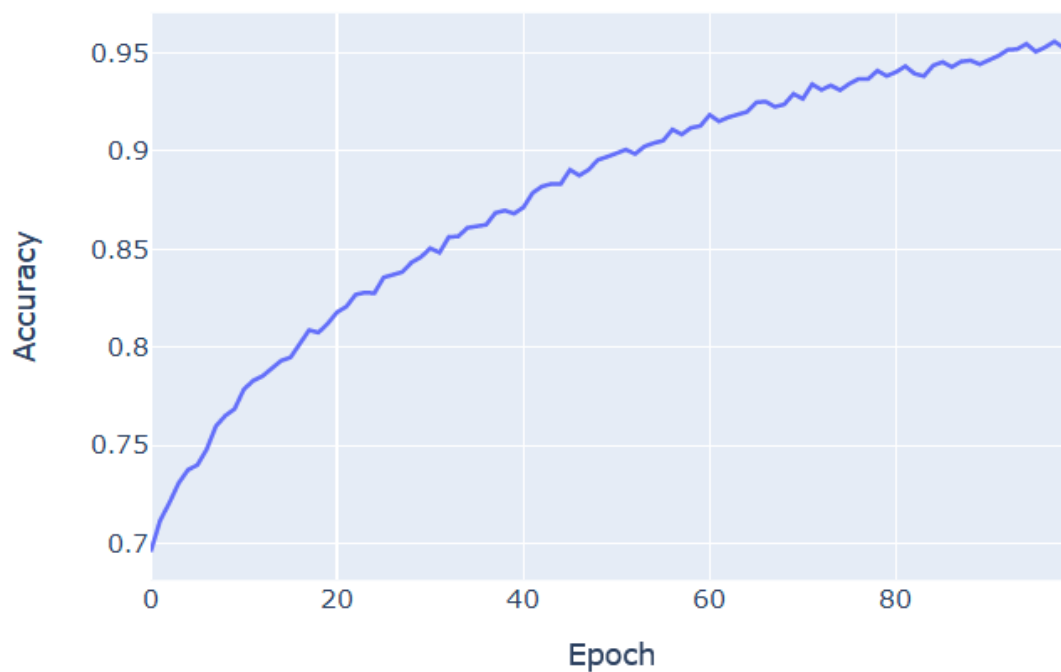
- `Working_singletask_fastai_v2_custom_implementation.ipynb`

Finálna a plne funkčná verzia multitask modelu aj s vyhodnotením napísaná v Kerase.

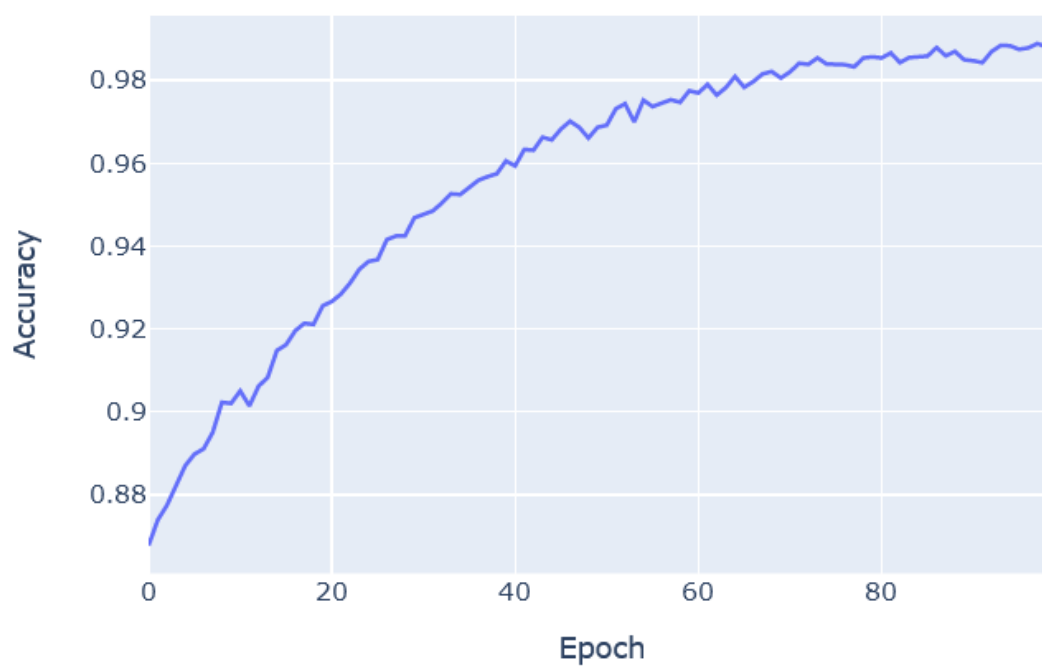
Záver

Dosiahnuté výsledky na trénovacej vzorke o veľkosti 11615 obrázkov. Tu bola použitá aj validačná sada o veľkosti 4978 obrázkov.

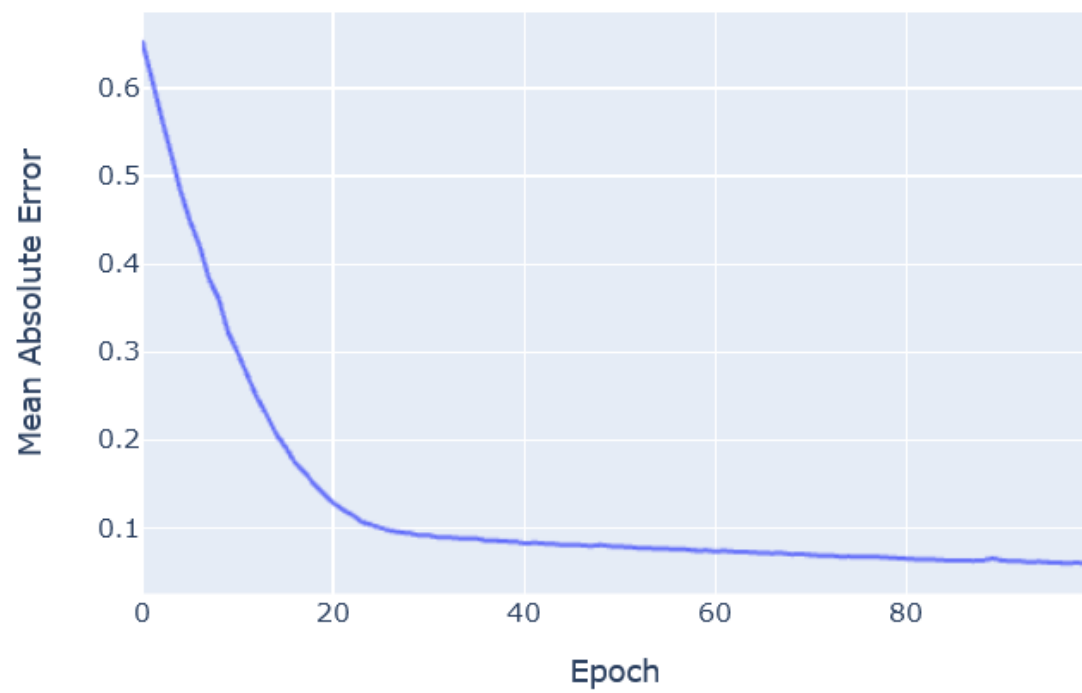
Accuracy for race feature



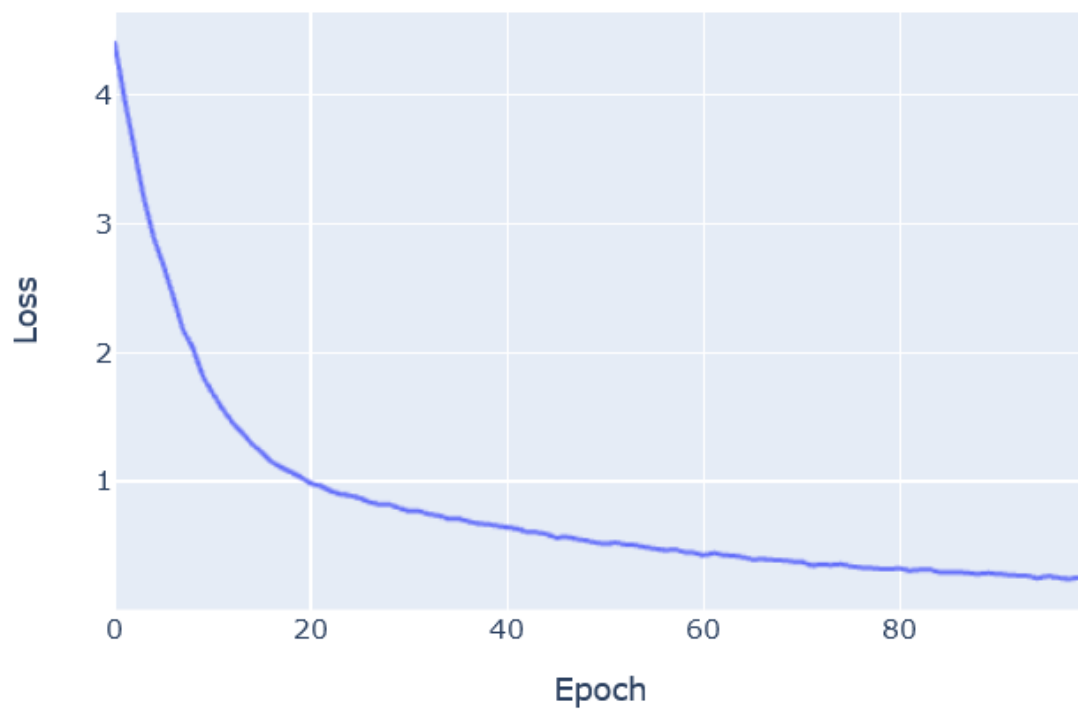
Accuracy for gender feature



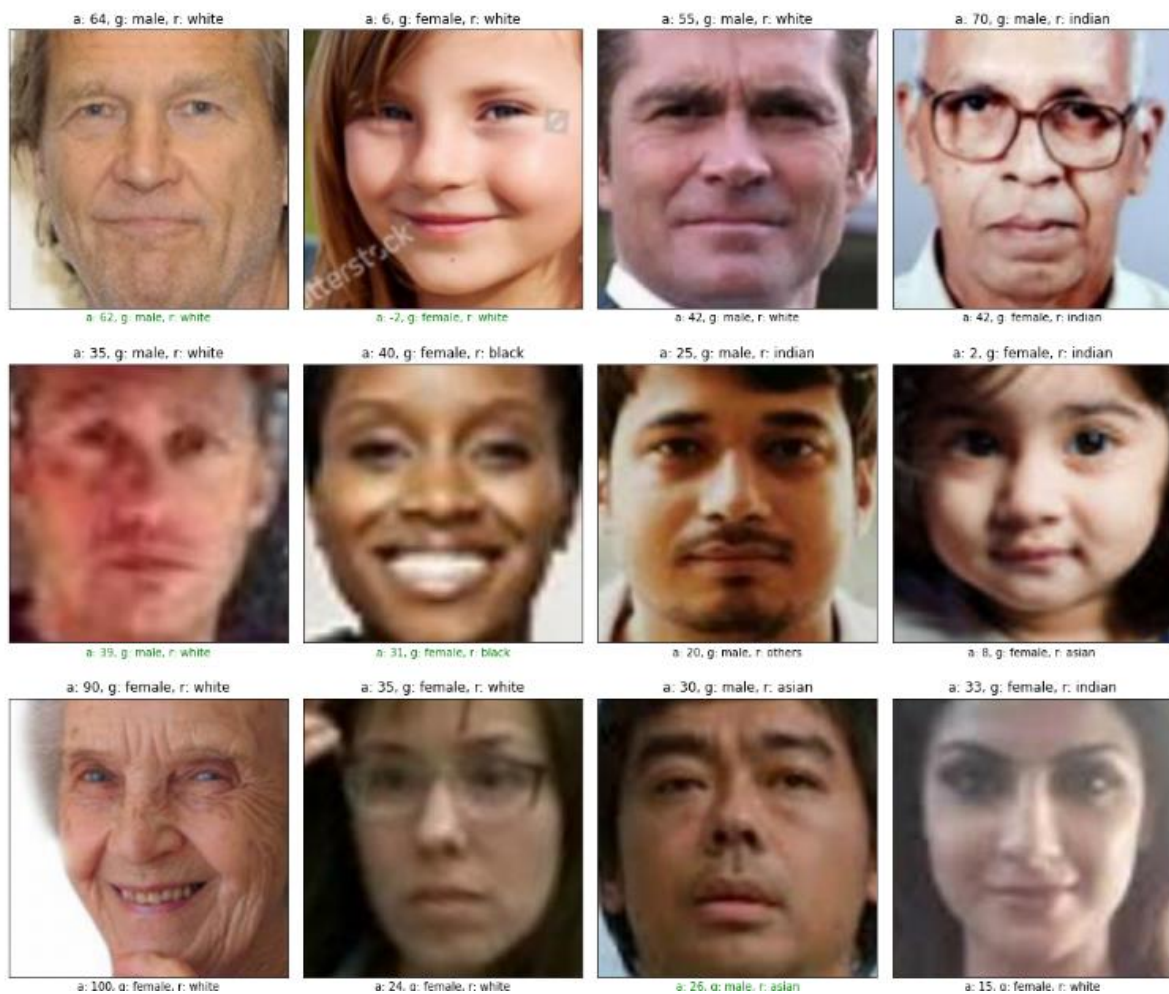
Mean Absolute Error for age feature



Overall loss



Výsledky dosiahnuté na testovacej vzorke o veľkosti 7112 obrázkov.



	precision	recall	f1-score	support
white	0.82	0.88	0.85	3071
black	0.89	0.81	0.85	1358
asian	0.75	0.88	0.81	977
indian	0.74	0.70	0.72	1212
others	0.44	0.27	0.34	486
accuracy			0.79	7104
macro avg	0.73	0.71	0.71	7104
weighted avg	0.79	0.79	0.79	7104

	precision	recall	f1-score	support
male	0.91	0.89	0.90	3747
female	0.88	0.91	0.89	3357
accuracy			0.90	7104
macro avg	0.90	0.90	0.90	7104
weighted avg	0.90	0.90	0.90	7104

R2 score for age: 0.5272995017753195