VMMT3 Projekt
IMDB review

Vypracoval: Jakub Čieško

Obsah

Popis úlohy

Úlohou je naučiť neurónovú sieť identifikovať polaritu filmovej recenzie.

Popis dát a preprocessing

IMDB reviews je veľká databáza filmových recenzií, ktorá sa bežne používa na trénovanie modelov na určovanie pozitívneho a negatívneho sentimentu textu. Dataset je voľne prístupný v knižnici Tensorflow pomocou príkazu tf.keras.datasets.imdb_reviews.load_data. Pozostáva z 50 000 recenzií rozdelených na train (25 000) a test (25 000) množiny. Dataset je vyvážený: pomerné zastúpenie jednotlivých tried je rovnaké (50 %).

Dataset je predspracovaný, je v podobe matice čísiel, je tokenizovaný. Pred jeho načítaním však treba zvoliť veľkosť slovníka (v našom prípade použijeme 2048 slov), ktorá určí hranicu najfrekventovanejších slov, a všetky ostatné zmení za výplňový token. Jediná potrebná úprava je nastavenie dĺžky recenzií. Aby bola stála, je treba určiť jednotnú dĺžku a všetky recenzie na ňu zarovnať (napr. pomocou funkcie tf.keras.utils.pad_sequences). Nami nastavená dĺžka je 256 slov. Ukážku recenzie v tokenizovanej a dĺžkovo homogenizovanej podobe vidíme na obrázku 1. Ukážku v slovnej podobe na obrázku 2.

```
array([   0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
          0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
          0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
          0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
          0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
          0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
          0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
          0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
          0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
          0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
          0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
          0,    0,    0,    0,    0,    1,   14,   20,   47,  111,  439,
          2,   19,   12,   15,  166,   12,  216,  125,   40,    6,  364,
        352,  707, 1187,   39,  294,   11,   22,  396,   13,   28,    8,
        202,   12, 1109,   23,   94,    2,  151,  111,  211,  469,    4,
         20,   13,  258,  546, 1104,    2,   12,   16,   38,   78,   33,
        211,   15,   12,   16,    2,   63,   93,   12,    6,  253,  106,
         10,   10,   48,  335,  267,   18,    6,  364, 1242, 1179,   20,
         19,    6, 1009,    7, 1987,  189,    5,    6,    2,    7,    2,
          2,   95, 1719,    6,    2,    7,    2,    2,   49,  369,  120,
          5,   28,   49,  253,   10,   10,   13, 1041,   19,   85,  795,
         15,    4,  481,    9,   55,   78,  807,    9,  375,    8, 1167,
          8,  794,   76,    7,    4,   58,    5,    4,  816,    9,  243,
          7,   43,   50], dtype=int32)
```

Obr. 1: Ukážka tokenizovanej recenzie spolu s výplňou na stanovenú dĺžku.

```
REVIEW_POLARITY: 0 - Negative
...<Padding Tokens>... this movie has many problem ...<Padding Tokens>... with it
that makes it come off like a low budget class project from someone in film school
i have to give it credit on its ...<Padding Tokens>... though many times throughout
the movie i found myself laughing ...<Padding Tokens>... it was so bad at times that
it was ...<Padding Tokens>... which made it a fun watch br br if you're looking for
a low grade slasher movie with a twist of psychological horror and a
...<Padding Tokens>... of ...<Padding Tokens>... then pop a ...<Padding Tokens>...
of ...<Padding Tokens>... some friends over and have some fun br br i agree with other
comments that the sound is very bad dialog is next to impossible to follow much of the
time and the soundtrack is kind of just there
```

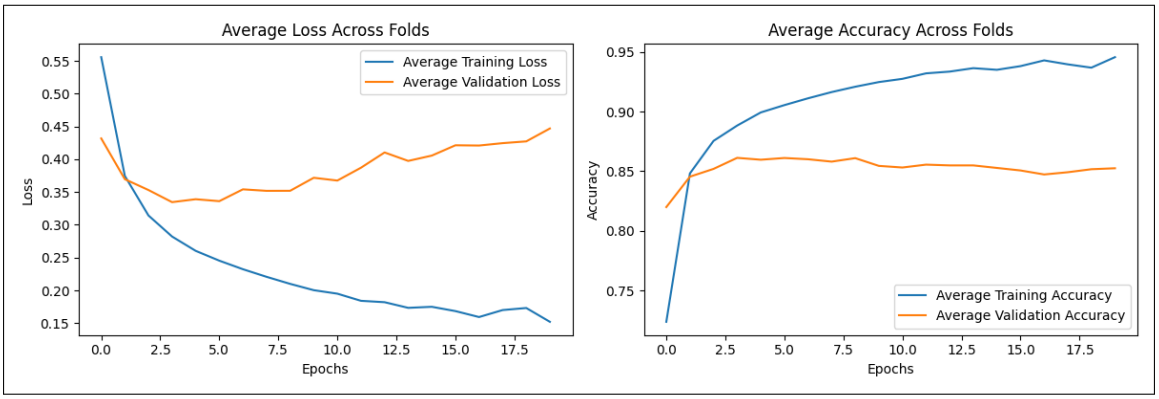
Obr. 2: Ukážka textovej recenzie z IMDB datasetu spolu s jej polaritou.

Model

Ako baseline si stanovíme jednoduchú rekurentnú sieť, ktorú sa finálnym modelom (zložitejšou rekurentnou sieťou) pokúsime prekonať. Pre lepšie vyhodnotenie presnosti modela použijeme pri trénovaní finálneho modela 5-fold cross validation.

Baseline

Baseline je jednoduchá rekurentná sieť, ktorej topológiu môžeme vidieť v tabuľke 1. Výsledky klasifikačných metrík a confusion matrix predikcií baseline modela na základe testovacieho datasetu (ktorý bol pred krížovou validáciou skrytý) môžeme vidieť v tabuľkách 2 a 3. Priebeh trénovania baseline modela (20 epoch, batch size 64, optimizér Adam(lr = 0.001) vidíme na obrázku 3. Ako vidíme, model po 3 epochách začína byť pretrénovaný.



Obr. 3: Priebeh trénovania baseline modela.

Name	Type	Shape	# Par.	Act.	Regul.	Padding
input_1	InputLayer	[(None, None)]	0	None	None	–
embedding_1	Embedding	(None, None, 16)	32768	None	None	–
lstm_1	LSTM	(None, 8)	800	tanh	None	–
dense_1	Dense	(None, 1)	9	sigmoid	None	–
Total par.:	33577	(131.16 KB)				
Trainable par.:	33577	(131.16 KB)				
Non-trainable par.:	0	(0.00 Byte)				

Tabuľka 1: Topológia baseline modela.

	Negative sent.	Positive sent.
Negative sent.	10898	1602
Positive sent.	1890	10610

Tabuľka 2: Confusion matrix baseline modela. Zvýraznená hlavná diagonála označuje správne klasifikácie.

	precision	recall	f1-score	support
Negative sent.	0.85	0.87	0.86	12500
Positive sent.	0.87	0.85	0.86	12500
accuracy			0.86	25000
macro avg	0.86	0.86	0.86	25000
weighted avg	0.86	0.86	0.86	25000

Tabuľka 3: Hodnota najbežnejších klasifikačných metrík baseline modela.

Trénovanie

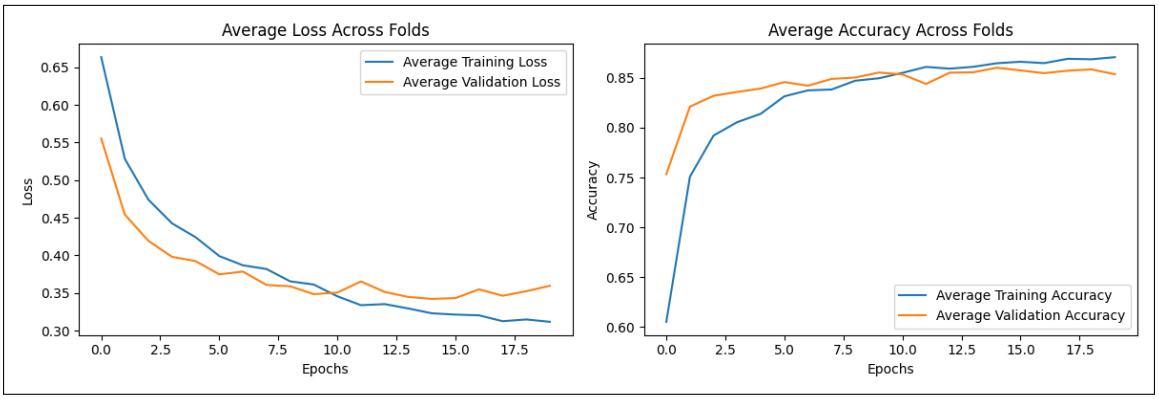
Finálny model bude zložitejšia rekurentná sieť zložená z vrstiev Embedding, Conv1D, Bidirectional LSTM, Attention a Dense. Conv1D do siete pridávame v snahe zachytiť opakujúce sa viacslovné motívy, n-gramy (kde n je definované veľkosťou jadra konvolúcie). Attention zas zaručí lepšie chápanie kontextu.

Aby sme zabránili preučeniu a ušetrili výpočtový čas zvýšime batch size na 128 a znížime počet epoch na 10.

Finálny model môžeme vidieť v tabuľke 4. Priebeh trénovania môžeme vidieť na obrázku 4. Model sme pomocou 5-fold cross validation trénovali 10 epoch, s rovnakým optimizátorom a jeho nastavením, pri batch size 128. Pri embedding layer sme nastavili mask_zero na True, a veľkosť jadra Conv1D vrstvy na 3.

Name	Type	Shape	#Par.	Act.	Reg.	Padding
input_1	InputLayer	[(None, None)]	0	None	None	–
embedding_1	Embedding	(None, None, 16)	32768	None	None	–
conv1d_1	Conv1D	(None, None, 4)	196	relu	L2=0.01	same
bidirectional_1	Bidirectional	(None, None, 4)	112	None	None	–
attention_1	Attention	(None, None, 4)	1	None	None	–
global_max_pool...	GlobalMaxPooling1D	(None, 4)	0	None	None	–
dropout_1	Dropout(0.5)	(None, 4)	0	None	None	–
dense_1	Dense	(None, 1)	5	sigmoid	None	–
Total par.:	33082	(129.23 KB)				
Trainable par.:	33082	(129.23 KB)				
Non-trainable par.:	0	(0.00 Byte)				

Tabuľka 4: Topológia modela.



Obr. 4: Priemerný priebeh trénovania finálneho modela pomocou 5-fold cross-validation.

Testovanie

Na vyhodnotenie modela sme použili testovací dataset, ktorý bol počas krížovej validácie skrytý. Confusion matrix predikcie testovacej množiny môžeme vidieť v tabuľke 5. Bežné hodnoty evaluácie modelov nájdeme v tabuľke 6.

	Negative sent.	Positive sent.
Negative sent.	10434	2066
Positive sent.	1271	11229

Tabuľka 5: Confusion matrix modela. Zvýraznená hlavná diagonála označuje správne klasifikácie.

	precision	recall	f1-score	support
Negative sent.	0.89	0.83	0.86	12500
Positive sent.	0.84	0.90	0.87	12500
accuracy			0.87	25000
macro avg	0.87	0.87	0.87	25000
weighted avg	0.87	0.87	0.87	25000

Tabuľka 6: Hodnota najbežnejších klasifikačných metrík modela.

Záver a porovnanie s baseline

Rozdieľ medzi jednoduchým baseline model a finálnym modelom nie je významný, napriek tomu však ide o malé zlepšenie (nárast o 1 % vo všetkých metrikách je v porovnaní s pôvodným výsledkom nárast o 1,16 %). Takýto výsledok je však pomerne očakávateľný, nakoľko je bežne uvádzaný v rôznych zdrojoch, kde sa IMDB dataset spracúva pomocou RNN a nie napr. transformermi (napr. Chollet 2022 s. 329, 332).

V tabuľke 7 vidíme rozdiely medzi confusion matrix finálneho modela a baseline modela. Negatívne hodnoty nesprávnych klasifikácií označujú tie predikcie, v ktorých sa finálny model mýli menej, a naopak pozitívne tie, v ktorých chybuje viac.

Hodnoty správnych klasifikácií – vyznačená hlavná diagonála – sa interpretujú opačne; pozitívny rozdiel znamená, že výsledný model určil danú triedu o daný počet klasifikácií častejšie než baseline, negatívne číslo zas znamená zhoršenie finálneho modela.

Vidíme, že finálny model je horší v odhalovaní negatívnych recenzií, ale za to je lepší v odhalovaní pozitívnych recenzií.

	Negative sent.	Positive sent.
Negative sent.	-464	464
Positive sent.	-619	619

Tabuľka 7: Rozdiel confusion matrix modela a baselinemodela.

Za poznámku ešte stojí to, že finálny model ukážku klasifikácie recenzie z obrázka 2 klasifikoval správne ako negatívnu recenziu.

Lepšie výsledky by sme mohli dostať napríklad pri použití iných, externých, už natrénovaných modelov (transfer learning), ktoré dokážu vyextrahovať z pôvodných dát viac informácií, potom by sme však trénovali skôr už len klasifikátor než celú sieť. Ďalším spôsobom, ako zrejme môžeme zlepšiť výsledky, je použitie vektorových embeddingov.