Review rating task
-SII-
Alexandru Cremeneanu

Am considerat acest task o problema de clasificare cu 5 clase, fiecare clasa reprezentand un rating.

Setul de date a fost impartit în 67% date pentru train și 33% date pentru test. Pentru validare, am folosit 10% din datele de antrenare.

Singurul pas de preprocesare al datelor a fost inlocuirea majusculelor cu litere mici, apoi acestea au fost tokenizate cu un tokenizer dintr-un model de distilbert (racai/distilbert-base-romanian-uncased). Fiecare propozitie a fost tokenizata pana la o lungime maxima de 512.
Pentru output am folosit one-hot encoding.

Pentru a rezolva problema claselor nebalansate am incercat mai multe variante:

- Reducerea de exemple care aveau rating 5 pentru a le aduce la acelasi nivel cu cele de rating 4
- Adaugarea de class_weights în timpul antrenarii modelelor
- Ambele variante de mai sus

Am decis să aleg varianta a 2-a, doar adaugarea de weight-uri pentru balansarea setului de date.

Pentru a alege hiperparametrii, am realizat un gridsearch pe:

- Embedding size
- Dropout rate
- LSTM size
- Learning rate
- Dense layer size

De asemenea, au fost incercate mai multe tipuri de modele folosind:

- LSTM
- Bidirectional LSTM
- GRU
- Bidirectional GRU

Arhitectura finala a modelului este urmatoarea:

- Embedding 128
- Bidirectional LSTM 20 merge_mode = sum
- Dropout 0.8

- Dense 16
- Dropout 0.8
- Dense output softmax
- Adam optimizer lr 2e-4
- Categorical Crossentropy loss

Rezultatele pentru cel mai bun model au fost de 0.71 weighted f1 score pe setul de testare (33% din setul de date initial)

Link notebook colab:
https://colab.research.google.com/drive/1_6W8VOjZ9_sDWjlbYLM9lbAvwCLed-df?usp=sharing

Link Github:
https://github.com/acremeneanu/SII_task