BERT. Классификация пар

Препроцессинг

В рамках предобработки текстовых данных были использовано два подхода для сравнения:

- использование такой же предобрабоки, как для базовых моделей
- приведение к нижнему регистру

После проведения экспериментов было выявлено улучшение на 0.01 по f-score при полной предобработке с токенезацией и лемматизацией. В итоге выбран был способ предобработки путем приведения текстов к нижнему регистру, т.к время обучения и предсказывания модели в этом случае увеличивается в 10 раз.

Далее, в связи с ограниченным количеством токенов, которые можно подать в модель (512), признаки ограничивались следующим образом: 5 (бренд), 20 (название), 20 (описание), 100 (характеристики). Полученные признаки конкатенировались для каждого товара в паре, а затем пара товаров объединялась следующим образом: [CLS] sequence1 [SEP] sequence2 [SEP], где sequence1 - вся информация о товаре 1, sequence2 - вся информация о товаре 2, [CLS] - начальный токен необходимый для модели, [SEP] - токен разделитель необходимый для модели. Также был протестирован подход постановки токена [SEP] между каждым признаком: [CLS] title1 [SEP] brand1[SEP] description1 [SEP] specifications1 [SEP] title2 [SEP] brand2 [SEP] description2 [SEP] specifications2 [SEP], но данный подход показал результаты хуже.

В качестве токенайзера текстов и модели-энкодера был использован bert-base-multilingual-uncased.

На выходе данная архитектура возвращает вектор размерности (768,), которые потом используется для классификации.

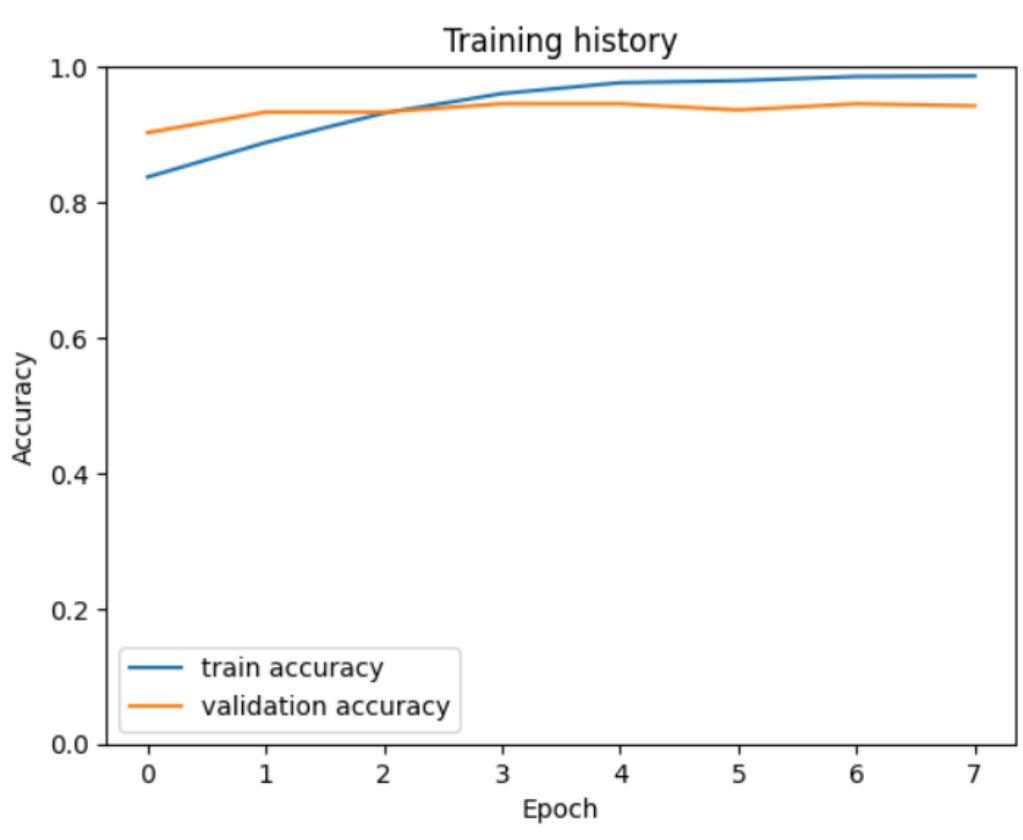
Эксперименты

Для обучения исользовались следующие параметры

name	best_val	vals_to_test
Optimizer	Adam	
Initial learning rate	1e-5	[5e-5,3e-5,2e-5,1e-5]
loss function	CrossEntropyLoss	
batch_size	16	[16,32,64]
EPOCHS	20	[3,7,10,20]

В качестве экспериментов над параметрами были исследованы различные значения, приведенные таблице, значения при которых удалось достичь лучших показателей расположены в столбце best_val.

Итоговое значение для количества эпох было выбрано - 20, но в случае если на протяжении 5 эпох улучшения на тестовой выборке нет, то обучение останавливается.



Также в ходе экспериментов выяснилось, что признак, содержащий характеристики товара привносил шум и ухудшал качество модели, поэтому этот признак в модель не вошел. Остальные признаки подавались в следующем порядке: title+brand+description.

Результаты

В результате проведения экспериментов были получены следующие результаты, которые превышают значения полученные с помощью базовых моделей:

	precision	recall	f1-score	support
0	0.96	0.97	0.97	305
1	0.90	0.90	0.90	106
accuracy			0.95	411
macro avg	0.93	0.93	0.93	411
weighted avg	0.95	0.95	0.95	411