Report

Search Summary

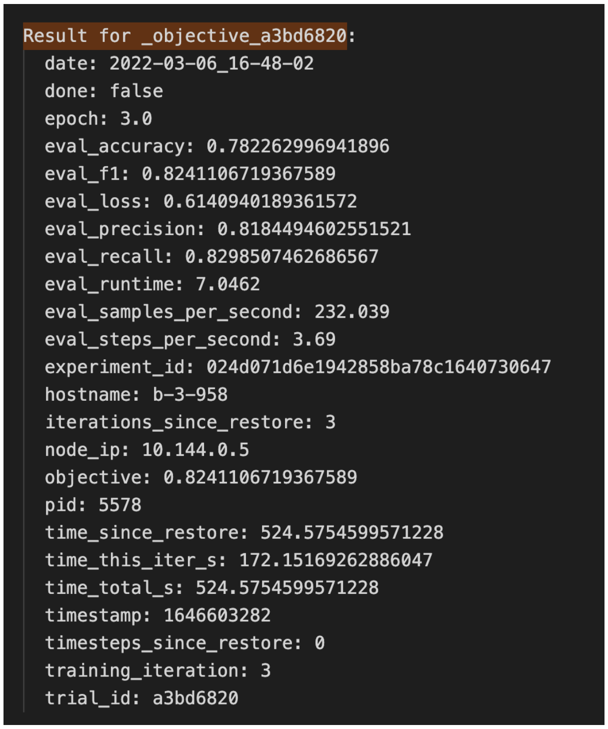
Trial #	Trial Name	Optimize Objective	Learning Rate	eval_accuracy	eval_f1	eval_precision	eval_recall	eval_loss	Screenshot
1	_objective_f847fce2	eval_f1	2.49816e-05	0.6147	0.7614	0.6147	1.0	0.6661	
2	_objective_a3bd6820	eval_f1	1.23233e-05	0.7823	0.8241	0.8184	0.8299	0.6141	
3	_objective_f8679534	eval_f1	4.80286e-05	0.6147	0.7614	0.6147	1.0	0.6634	
4	_objective_31405ed0	eval_f1	3.92798e-05	0.6147	0.7614	0.6147	1.0	0.6667	
5	_objective_6ac9adb8	eval_f1	3.39463e-05	0.6147	0.7614	0.6147	1.0	0.6667	

Report the evaluation metrics and tuned hyperparameters of your best run. Were there any other models that had higher loss but better evaluation accuracy or f1 score? Did the objective value vary a lot across runs?

Answer to
Report the evaluation metrics and tuned hyperparameters of your best run.

Best Run

> ### Evaluation Metric
>
> - eval_accuracy
>   0.7823
> - eval_f1
>   0.8241
> - eval_precision
>   0.8184
> - eval_recall
>   0.8299
> - eval_loss
>   0.6141



```
Result for _objective_a3bd6820:
  date: 2022-03-06_16-48-02
  done: false
  epoch: 3.0
  eval_accuracy: 0.782262996941896
  eval_f1: 0.8241106719367589
  eval_loss: 0.6140940189361572
  eval_precision: 0.8184494602551521
  eval_recall: 0.8298507462686567
  eval_runtime: 7.0462
  eval_samples_per_second: 232.039
  eval_steps_per_second: 3.69
  experiment_id: 024d071d6e1942858ba78c1640730647
  hostname: b-3-958
  iterations_since_restore: 3
  node_ip: 10.144.0.5
  objective: 0.8241106719367589
  pid: 5578
  time_since_restore: 524.5754599571228
  time_this_iter_s: 172.15169262886047
  time_total_s: 524.5754599571228
  timestamp: 1646603282
  timesteps_since_restore: 0
  training_iteration: 3
  trial_id: a3bd6820
```

Answer to
Were there any other models that had higher loss but better evaluation accuracy or f1 score?

> All four other runs, as we can see in the table at the beginning, have lower `eval_accuracy` and higher `eval_loss` than the best run. However, it is worth noting that all four other runs feature a higher `eval_recall` than the best run (all four being 1.0).

Answer to
Did the objective value vary a lot across runs?

> The objective value (in my case is `eval_f1`) vary little in other four trials (excluding the best run).

My Observation

> From the five trials, maybe the best way to fine-tune Roberta on BoolQ is to use a relatively smaller learning rate (something around 1.2e-5). When learning rate is high (e.g., higher than 2.5e-5), it seems that the model will start to push `eval_recall` to 1.0, which may be signaling that the model is always guessing “positive”.