A Re-implementation of Baselines

While Zhou et al. (2020) provided official implementation of the proposed method, we have several concerns regarding its reproducibility and experimental setup. The original implementation implies running evaluation iteration-wise (i.e., the best performing model is selected once in n iterations).

In this case, several iterations n between validation epochs could be seen as an additional hyperparameter, which has a large impact on the resulting performance. Although, Zhou et al. (2020) did not include the best performing value so that the results could be easily reproduced.

Because of this, we decided to re-implement PABEE strictly following the original paper and training details, while evaluating the model after every epoch, instead of evaluating after n training steps, which we believe is a fairer way to compare methods.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 4.
   (b) Did you describe the limitations of your work? [Yes] See Section 5
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] We proposed a method that allows us to perform an early exit from an ALBERT model. We believe that there are no any potential negative societal impacts of our work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We ran all experiments on openly available GLUE benchmark datasets and provide source code for our paper. We also provide WandB configs for hyperparameter search and best model evaluation on dev split of the data so that the results could be easily reproduced.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We conducted all experiments on default splits of the GLUE benchmarking datasets. Hyperparameter ranges are available in Table 4.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We reported mean and std values across different seeds for the ablation study (see Table 1), while for the main experiment, we decided to report only median values across runs for consistency with recent works. During the exploration of the proposed method in Sections 4.3, 4.5, we also reported std values across runs as well as mean values.
   (d) Did you include the total amount of computing and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] We did not include this information due to NDA concerns.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 4.1
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]