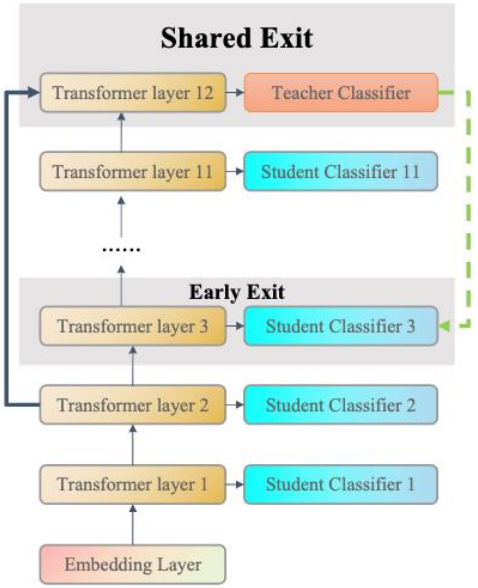
Accelerating BERT Inference with GPU-Efficient Exit Prediction

Lei LI, Chengyu WANG, Minghui QIU,
Cen CHEN, Ming GAO, Aoying ZHOU

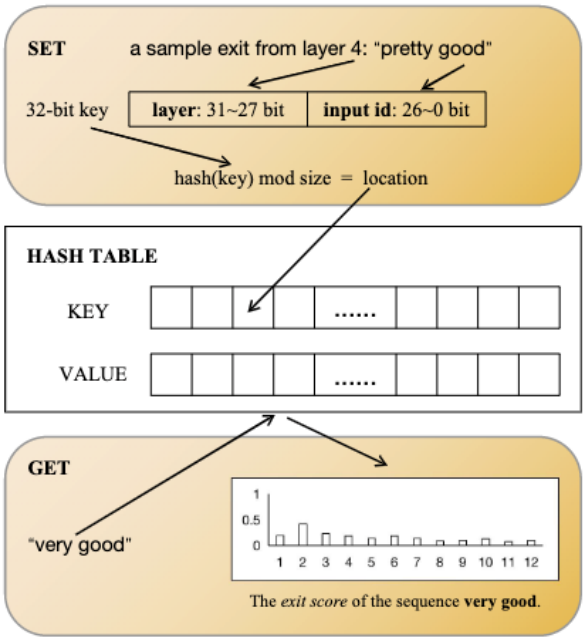
Frontiers of Computer Science, DOI: 10.1007/s11704-022-2341-9

Problems & Ideas

- Problems of FastBERT:
  - The teacher classifier is not knowledgeable enough.
  - The batch size shrinkage and the redundant computation of student classifiers.
- Ideas: (1) Shared Exit Loss: making the teacher classifier more knowledgeable by feeding diverse Transformer outputs to it. (2) Exit Layer Prediction: utilizing a GPU hash table to handle the token-level exit layer distribution and sorting test samples by predicted exit layers.



Shared Exit Loss



Exit Layer Prediction

Main Contributions

- Contributions:
  - We propose GEEP including two significant components Shared Exit Loss (SEL) and Exit Layer Prediction (ELP);
  - Experimental results show that GEEP outperforms FastBERT and other Early-Exit models in most situations; the ablation study proves the effectiveness of SEL and ELP.