Cornell Bowers C·IS College of Computing and Information Science

DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter

Victor SANH, Lysandre DEBUT, Julien CHAUMOND, Thomas WOLF
Hugging Face
{victor,lysandre,julien,thomas}@huggingface.co

Presented by Xinlin Wu, Bozhou Jin, Chengyi Jiang

Paper Summary & Key Contributions

DistilBERT shrinks BERT by 40 % yet preserves 97 % of its accuracy and runs 60 % faster by applying knowledge distillation during pre-training. A triple loss—masked-LM, soft-target distillation, and cosine alignment—lets the smaller 6-layer model inherit the teacher's linguistic knowledge, making it cheap to train and practical for on-device NLP.

Introduction / Background / Motivation

- **Problem Addressed**

  Large-scale language models such as BERT-base (≈ 110 M parameters) incur high latency and memory footprints that block real-time, on-device NLP deployment .

- **Project Goal / Hypothesis**

  Reproduce DistilBERT using smaller training datasets, and verify that it has performance comparable to the teacher BERT model.

- **Target Result**

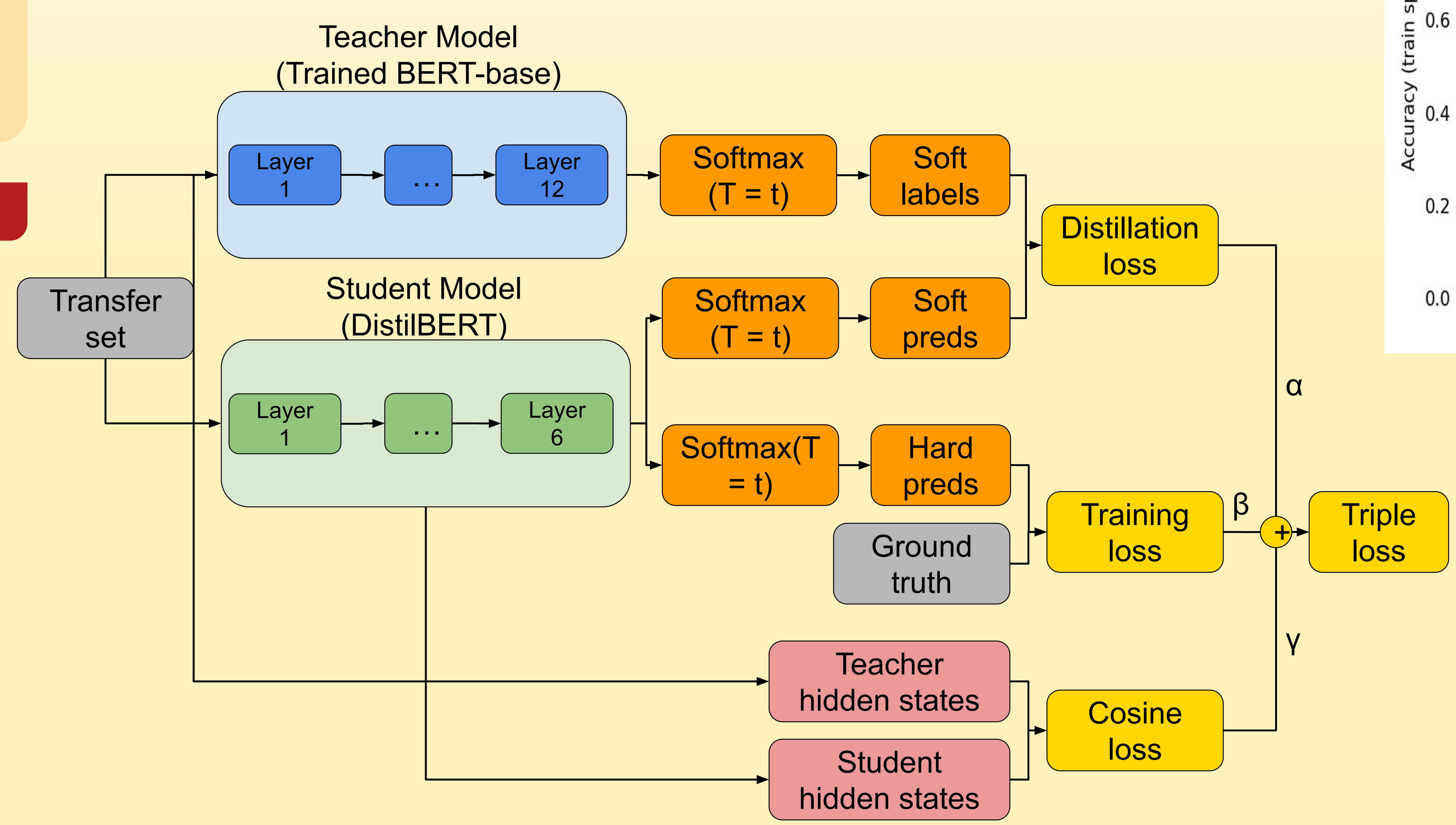
  Reproduce the macro GLUE score reported by Sanh et al.—DistilBERT 77.0 vs. BERT-base 79.5 (Table 1 of the paper) , by evaluating with data sampled from the provided basic datasets.

- **Context & Motivation**

  If the compressed model matches within ~3 % of BERT while being **40 % smaller and 60 % faster** , practitioners gain a sustainable, edge-friendly alternative—lower energy cost, faster iteration, and wider accessibility.

Methodology

DistilBERT is trained from the BERT-base model through knowledge distillation. It retains the overall architecture of BERT-base, but reduces the number of Transformer encoder layers from 12 to 6, and removes the token-type embeddings and the pooler.

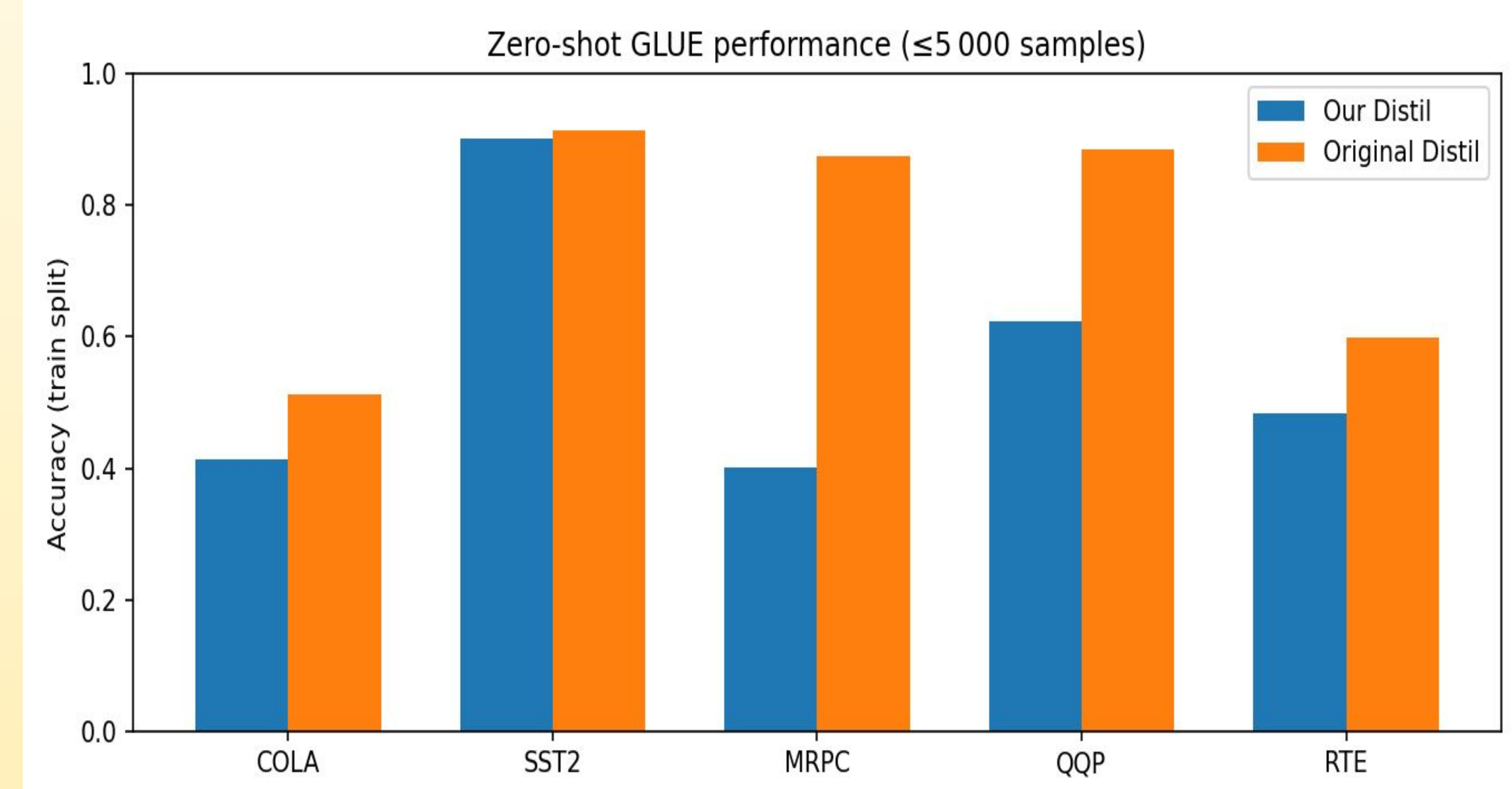


Our modification

- **Dataset:** Instead of pretraining on the full concatenation of English Wikipedia and the Toronto Book Corpus, we use a significantly smaller dataset. Specifically, we merge the training and validation sets of SST-2 from the GLUE benchmark, and split the combined data into 80% for training, 10% for validation, and 10% for testing.
- **Weight Initialization**: Unlike the original DistilBERT which initializes the student model using the teacher's weights, our student model is initialized with random weights.

Tools and Hardware: We implement our approach using the Hugging Face Transformers library and train the model on a laptop equipped with an NVIDIA RTX 4070 GPU.

Results



Conclusion

1. Training with only 1 epoch on SST2 is sufficient for DistilBERT to achieve high performance (~90%) on tasks within the dataset.
2. This is comparable with the result in the original distilBERT paper (91.3%).
3. Additional training is required for zero-shot performance on external datasets (especially MRPC and QQP)

References

- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108. https://arxiv.org/abs/1910.01108
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.
- 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 7th International Conference on Learning Representations (ICLR 2019)*. https://openreview.net/forum?id=rJ4km2R5t7