Cornell Bowers C·IS
College of Computing and Information Science

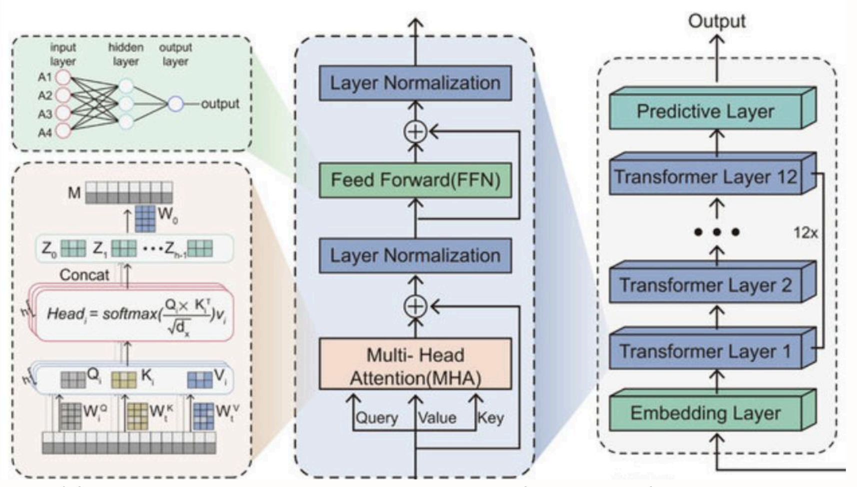
TinyDistilBERT, a distilled version of BERT: a lot smaller, faster, cheaper and lighter

Haofan Wang, Eric Cao, Ian Chen, Solomon Lee, Shohaib Shah

Cornell University

Problem

Ways to address **computational burdens** and **speed limitations** of LLMs.



source: https://www.researchgate.net/figure/Schematic-diagram-of-BERT-BASE-and-DistilBERT-model-architecture_fig1_382939584

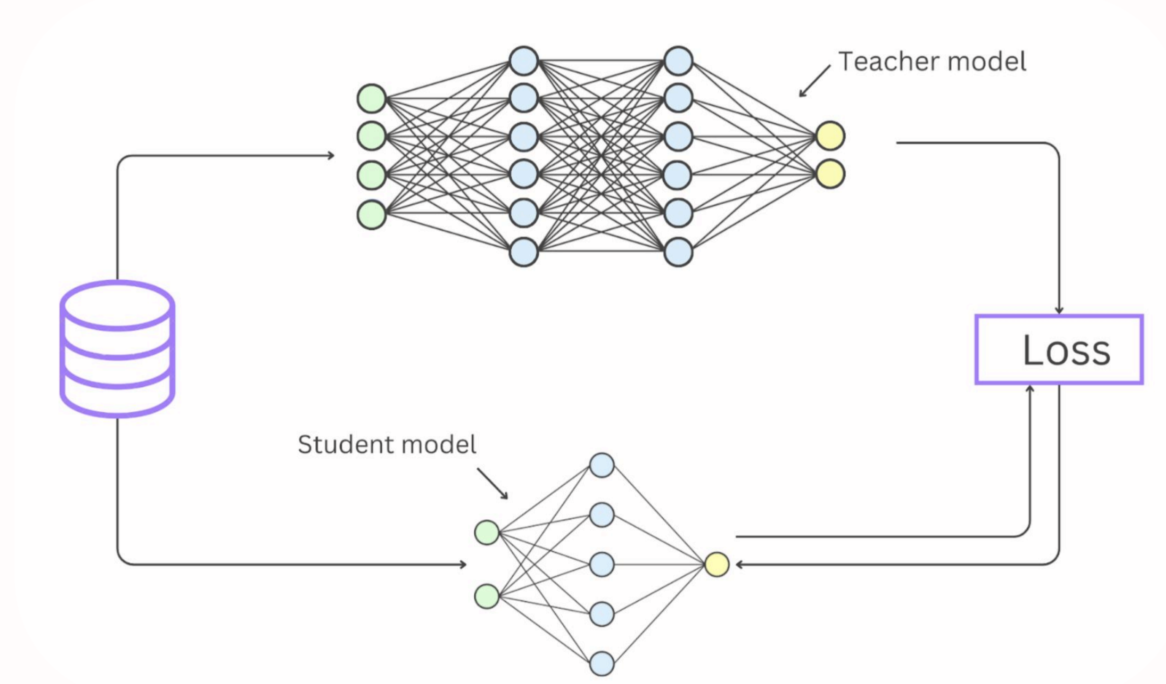
Current practice is having many layers, resulting in:

computation costs memory demands poor inference speed

Idea

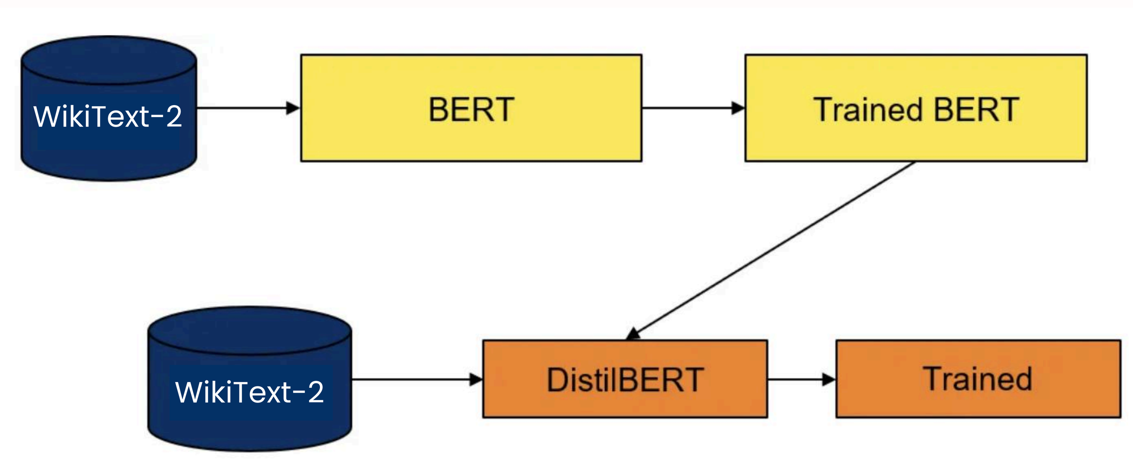
Tokenized Text

Train a **smaller**, simpler student model by replicating the **knowledge** from a **larger**, teacher model



source: https://www.researchgate.net/figure/Schematic-diagram-of-BERT-BASE-and-DistilBERT-model-architecture_fig1_382939584

Teacher trains student network by feeding it soft probabilities, helping the student **learn** and **replicate** the teacher's behavior.
BERT will serve as the teacher to train a student, DistilBERT.

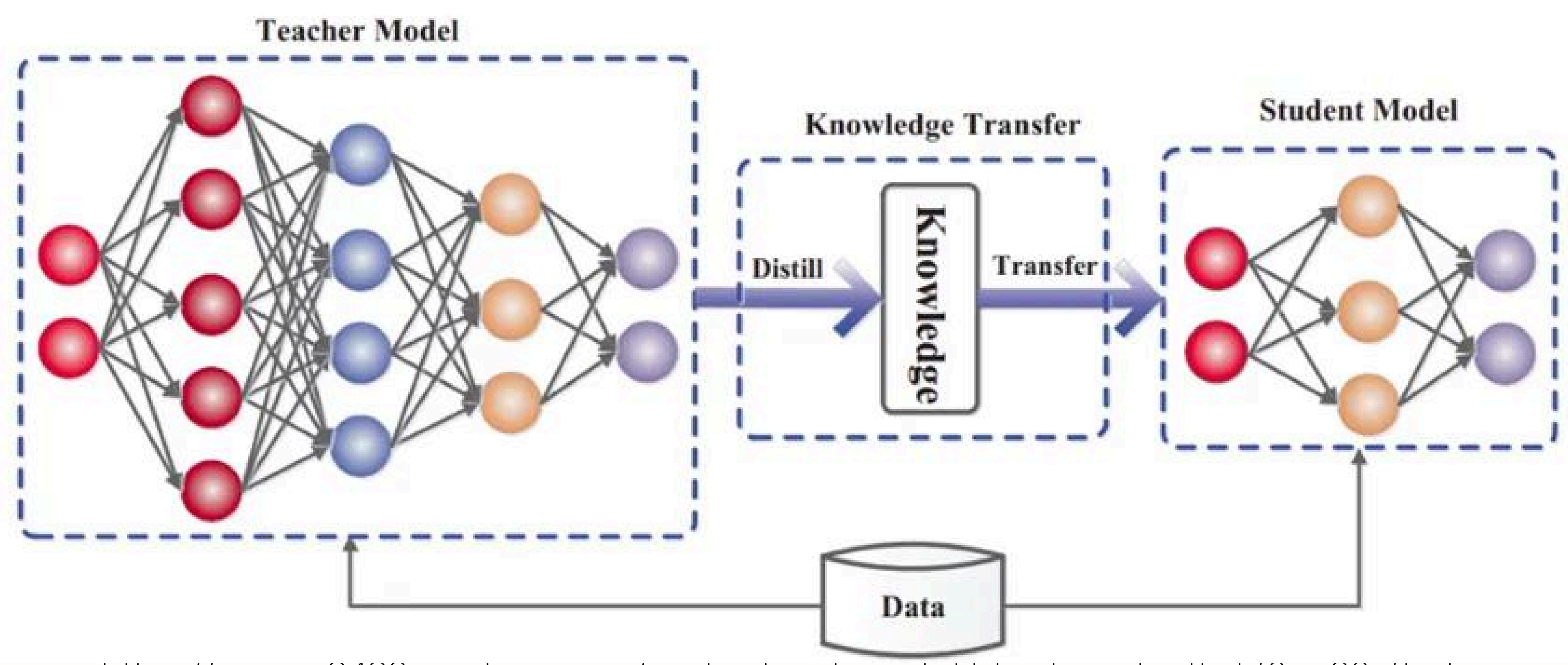


Our aim is to reproduce the distillation process of the original DistilBERT model. Whose result was a model that was 40% smaller, 60% faster, and retained 97% of BERT's capabilities.

Methodology

Pretraining: Distilled from BERT teacher, performed pretraining using wikitext-2 dataset (1/10 size of original training corpus)

Loss function: KL-loss + MLM cross-entropy + cosine loss



source: https://arunm8489.medium.com/understanding-distil-bert-in-depth-5f2ca92cf1ed

Finetuning: TinyDistilBert was finetuned separately on SQuAD (QA), IMDB (Binary Sentiment Classfication), and GLUE.

SQuAD
The Stanford Question Answering Dataset

source: https://medium.com/syncedreview/acl-best-paper-tricky-stanford-dataset-adds-questions-that-dont-have-answers-d7d95f4369df

source: https://www.kaggle.com/datasets/hijest/genre-classification-dataset-imdb

GLUE

source: https://gluebenchmark.com/

Inference Speed: Measure end-to-end inference time on CPU (single core) and GPU for a batch size of 1

Design Choices / Modifications

1. Used WikiText-2 as our training corpus due to original training corpus size being too large for our GPUs to train on in time
2. Separate python notebooks.
   a. We had 6 python notebooks: 1 pretraining, 1 GLUE, 1 IMDb, 2 SQUAD, 1 GLUE inference
3. Due to smaller training corpus, we had 4-5 fine-tuning epochs for each GLUE task as some GLUE tasks were returning a score of 0.

Results

Table 1: **TINYDistilBERT retains large amount of BERT performance (70%).**

Model	Score	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST-2	STS-B	WNLI
ELMo	68.7	44.1	68.6	76.6	71.1	86.2	53.4	91.5	70.4	56.3
BERT-base	79.5	56.3	86.7	88.6	91.8	89.6	69.3	92.7	89.0	53.5
DistilBERT	77.0	51.3	82.2	87.5	89.2	88.5	59.9	91.3	86.9	56.3
TinyDistillBERT	54.9	12.8	60.3	80.5	60.2	72.8	53.0	79.2	18.1	56.3

source of ELMo, BERT-base, DistilBERT performance: https://arxiv.org/pdf/1910.01108

Table 2: **TinyDistilBERT yields to comparable performance on IMDb and less on SQuAD**

Model	IMDb (acc.)	SQuAD (EM/F1)
BERT-base	93.46	81.2/88.5
DistilBERT	92.82	77.7/85.8
DistilBERT (D)	-	79.1/86.9
TinyDistilBERT	86.19	10.36/18.58
TinyDistilBERT (D)	-	10.98/19.62

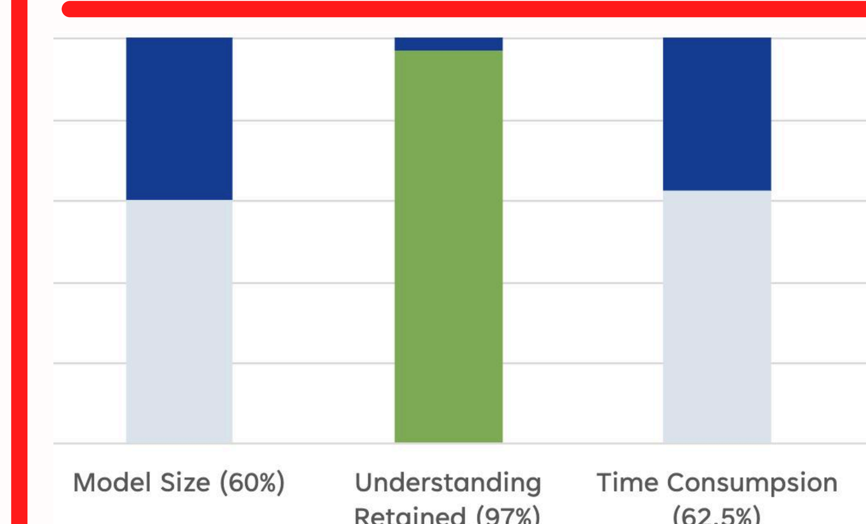
source of BERT-base, DistilBERT, DistilBERT (D) performance: https://arxiv.org/pdf/1910.01108

Table 3: **TinyDistilBERT is significantly smaller while being constantly faster.**

Model	# param. (Millions)	Inf. time (seconds)
TINYDistilBERT	66	215.79
DistilBERT	66	209.1
BERT-base	110	423.65
ELMo	180	895

source of ELMo performance: https://arxiv.org/pdf/1910.01108

Conclusion



- Significantly compressed the model
- Greater efficiency
- Excels on some semantic and sentiment tasks
- Struggles on syntax-sensitive, semantic-similarity, and QA benchmarks

References

[1] https://arxiv.org/abs/1910.01108
[2] https://paperswithcode.com/dataset/wikitext-2
[3] https://gluebenchmark.com/
[4] https://www.kaggle.com/datasets/hijest/genre-classification-dataset-imdb
[5] https://rajpurkar.github.io/SQuAD-explorer/