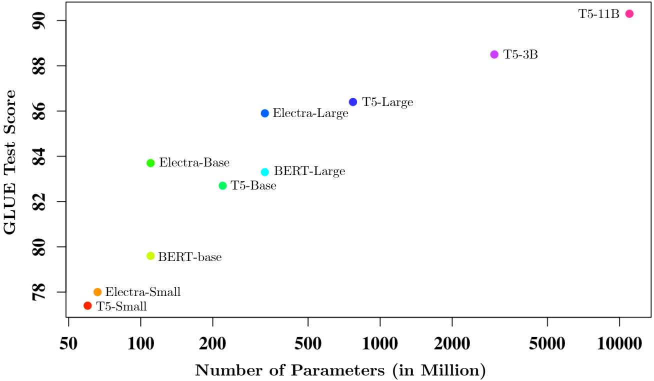
HomoDistil: Homotopic Task-Agnostic Distillation of Pre-trained Transformers

Chen Liang[*], Haoming Jiang[⋆], Zheng Li[⋆], Xianfeng Tang[⋆], Bin Yin[⋆], Tuo Zhao[*]

[*]Georgia Institute of Technology, [⋆]Amazon

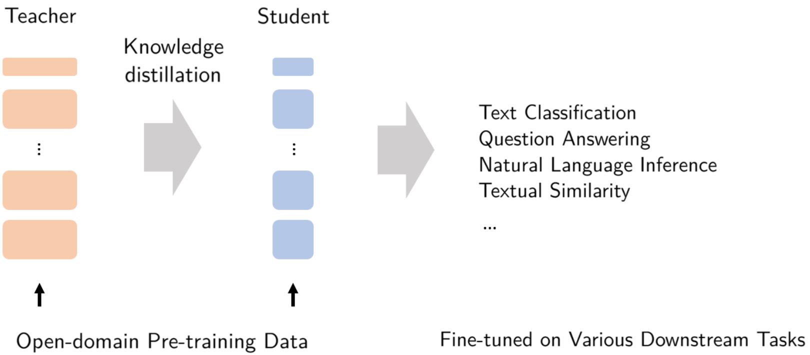
Apr. 1 2023

Growing Sizes of Language Models



This poses great challenges for model deployment on devices with latency requirements and memory constraints.

Task-Agnostic Distillation



Teacher

Knowledge distillation

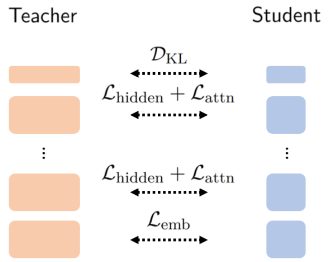
Student

Text Classification
Question Answering
Natural Language Inference
Textual Similarity
...

Open-domain Pre-training Data

Fine-tuned on Various Downstream Tasks

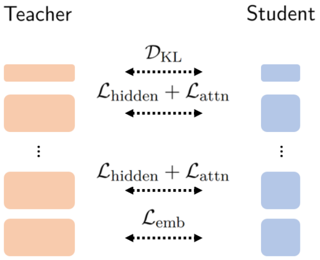
Layerwise Distillation


Teacher
Student
$\mathcal{D}_{\mathrm{KL}}$
$\mathcal{L}_{\mathrm{hidden}} + \mathcal{L}_{\mathrm{attn}}$
⋮
⋮
$\mathcal{L}_{\mathrm{hidden}} + \mathcal{L}_{\mathrm{attn}}$
$\mathcal{L}_{\mathrm{emb}}$

Objective:
$$\begin{aligned}\min_{\theta_{\mathrm{s}}} \mathcal{L}_{\mathrm{MLM}}(\theta_{\mathrm{s}}) \\ +\alpha_1 \mathcal{D}_{\mathrm{KL}}(\theta_s, \theta_t) \\ +\alpha_2 \mathcal{L}_{\mathrm{hidden}}(\theta_s, \theta_t) + \alpha_3 \mathcal{L}_{\mathrm{attn}}(\theta_s, \theta_t) + \alpha_4 \mathcal{L}_{\mathrm{emd}}(\theta_s, \theta_t).\end{aligned}$$

where θ_s: student model; θ_t: teacher model.

Layerwise Distillation


Teacher
Student
$\mathcal{D}_{\text{KL}}$
$\mathcal{L}_{\text{hidden}} + \mathcal{L}_{\text{attn}}$
⋮
⋮
$\mathcal{L}_{\text{hidden}} + \mathcal{L}_{\text{attn}}$
$\mathcal{L}_{\text{emb}}$

$$\mathcal{L}_{\text{hidden}}(\theta_s, \theta_t) = \sum_{k \in K} \texttt{MSE}(H_s^k, H_t^k W_{\text{hidden}}^k).$$

$$\mathcal{L}_{\text{attn}}(\theta_s, \theta_t) = \sum_{k \in K} \texttt{MSE}(A_s^k, A_t^k).$$

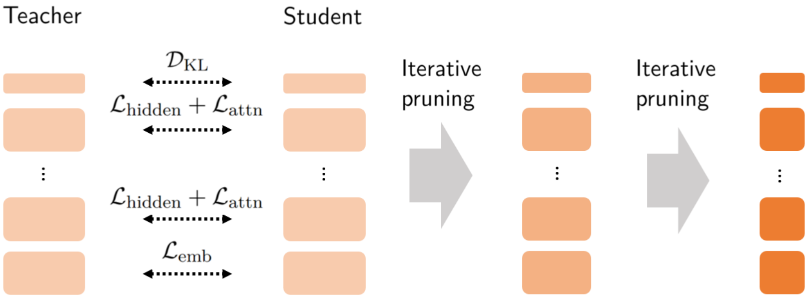
$$\mathcal{L}_{\text{emb}}(\theta_s, \theta_t) = \texttt{MSE}(E_s, E_t W_{\text{emb}}).$$

Large Teacher-Student Knowledge Gap

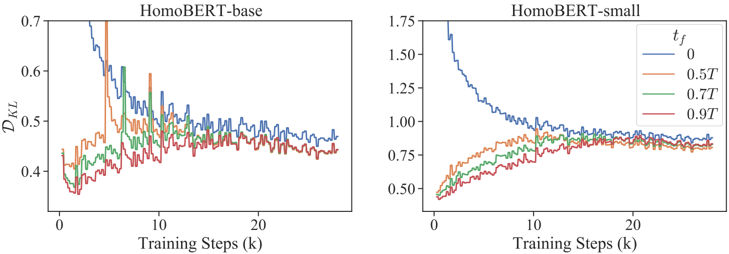
- There are large discrepancies between the student's and the teacher's layerwise representations.
- The student struggles to mimic the layerwise representations of the teacher.
- The student training favors reducing such large discrepancies over the training loss and underfits the training data.

HomoDistil: Maintain a Small Knowledge Gap

Initialize the student from the teacher and iteratively prune the student's neurons until the target width is reached.


Teacher
Student
$\mathcal{D}_{\mathrm{KL}}$
$\mathcal{L}_{\mathrm{hidden}} + \mathcal{L}_{\mathrm{attn}}$
$\mathcal{L}_{\mathrm{hidden}} + \mathcal{L}_{\mathrm{attn}}$
$\mathcal{L}_{\mathrm{emb}}$
Iterative pruning
Iterative pruning

HomoDistil: Maintain a Small Knowledge Gap



t_f: The number of iterations to achieve the target width.

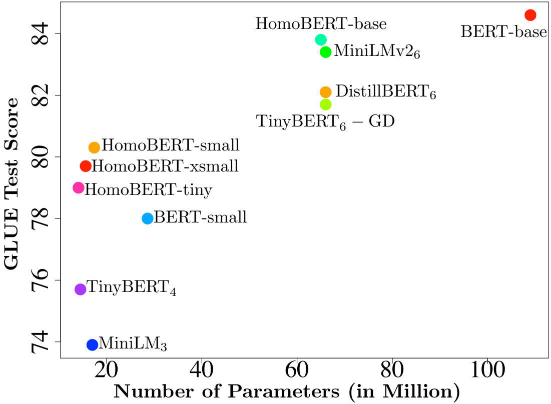
Experiment Settings

Student Architectures:

	Params (million)				
Model	Embedding	Backbone	Total	d^{hidn}	d^{ffn}
BERT-base (Teacher)	23.4	85.5	109	768	3072
HomoBERT-base	17.6	47.8	65	576	2304
HomoBERT-small	7.8	9.4	17.3	256	1024
HomoBERT-xsmall	7.3	8.3	15.6	240	960
HomoBERT-tiny	7.2	6.8	14.5	224	896

Distillation Dataset: Wikipedia + Bookcorpus.
Evaluation Dataset: GLUE benchmark.

Compare with Task-Agnostic Methods



DistilBERT (Sanh et al. 2019), TinyBERT (Jiao et al. 2020), MiniLMv2 (Wang et al., 2020).