Knowledge Distillation on General LM

by WangYC

Sep.30th 2022 @RUC

目录

1. 知识蒸馏基础知识

- 1.1 知识蒸馏涉及到的基本概念
  - 1.1.1 最早的蒸馏思想
  - 1.1.2 基本概念 teacher，student，temperature等
- 1.2 知识蒸馏研究framework
  - 1.2.1 研究思路：选定目标模型，缩小数据集（transfer set）
  - 1.2.2 评价指标

1.1 基本概念

- 最早的知识蒸馏思想:
  - 2015年Hinton等：将十个不同随机初始化DNN模型结果整合到一个相同结构的模型当中, 得到了比单独模型更优的结果

Distilling the Knowledge in a Neural Network

Geoffrey Hinton[*†] Google Inc. Mountain View geoffhinton@google.com

Oriol Vinyals[†] Google Inc. Mountain View vinyals@google.com

Jeff Dean Google Inc. Mountain View jeff@google.com

System	Test Frame Accuracy	WER
Baseline	58.9%	10.9%
10xEnsemble	61.1%	10.7%
Distilled Single model	60.8%	10.7%

- 如今知识蒸馏被广泛用于模型压缩

1.1 基本概念

- Teacher, Student(distilled model)
- hard / soft label & logits
- Temperature T

$$q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}$$

Training loss The student is trained with a distillation loss over the soft target probabilities of the teacher: $L_{ce} = \sum_i t_i * \log(s_i)$ where t_i (resp. s_i) is a probability estimated by the teacher (resp. the student). This objective results in a rich training signal by leveraging the full teacher distribution. Following Hinton et al. [2015] we used a *softmax-temperature*: $p_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$ where T controls the smoothness of the output distribution and z_i is the model score for the class i. The same temperature T is applied to the student and the teacher at training time, while at inference, T is set to 1 to recover a standard *softmax*.

The final training objective is a linear combination of the distillation loss L_{ce} with the supervised training loss, in our case the *masked language modeling* loss L_{mlm} [Devlin et al., 2018]. We found it beneficial to add a *cosine embedding* loss (L_{cos}) which will tend to align the directions of the student and teacher hidden states vectors.

1.2 研究framework

- 选定框架（Bert / GPT / GLM / ...）
- transfer set
- 目标：在尽可能保证效果不降的情况下缩小模型参数

目录

2. 以BERT蒸馏为例看传统蒸馏方法

- 2.1 尝试不同阶段的蒸馏
  - 2.1.1 微调蒸馏
  - 2.1.2 预训练蒸馏
  - 2.1.3 预训练 + 微调
- 2.2 尝试不同蒸馏对象
  - soft / hard label
  - logits
  - embedding + x
  - FFN
  - Attention / Q K V relationship

2. 以BERT蒸馏为例看传统蒸馏方法

方法	蒸馏次数	蒸馏阶段	中间通用特征							最后通用特征		任务特定特征				学生/教师相同约束
												预训练		微调		
			emb+	A-scores	A-probs	key	query	value	tf层	最后tf层	logits	MLM	NSP	soft	hard	
KD	1	微调												CE	CE	无
BERT-PKD	1	微调							MSE	MSE				KL	CE	隐层维度
DistilBERT	1	预训练								Cos	KL	CE				隐层维度
MiniLM	1	预训练			KL			KL								注意力头数
MiniLMv2	1	预训练				KL	KL	KL								无
TinyBERT	3	预训练/微调	MSE	MSE					MSE	MSE				CE		注意力头数
Mixbaseline (try)	3	预训练/微调	MSE	MSE	KL	KL	KL	KL	MSE	MSE	KL	CE		KL	CE	注意力头数

目录

3. 传统方法瓶颈&改进思路

- 3.1 教师与学生之间的gap问题
- 3.2 思路1: 继续寻找最优组合
  - 3.2.1 当前实验最优组
  - 3.2.2 实验新发现
- 3.3 思路2: 多教师方法
  - 3.3.1 多相同结构教师加权
  - 3.3.2 助教（迭代式）

思路1: 继续排列组合

- 3.2.1 原思路找到的最优解:
  - pre-train和finetune都将attention去掉，finetune将hard-label去掉

- 3.2.2 soft is all you need（

思路2: 多teacher

- 3.3.1 加权方法

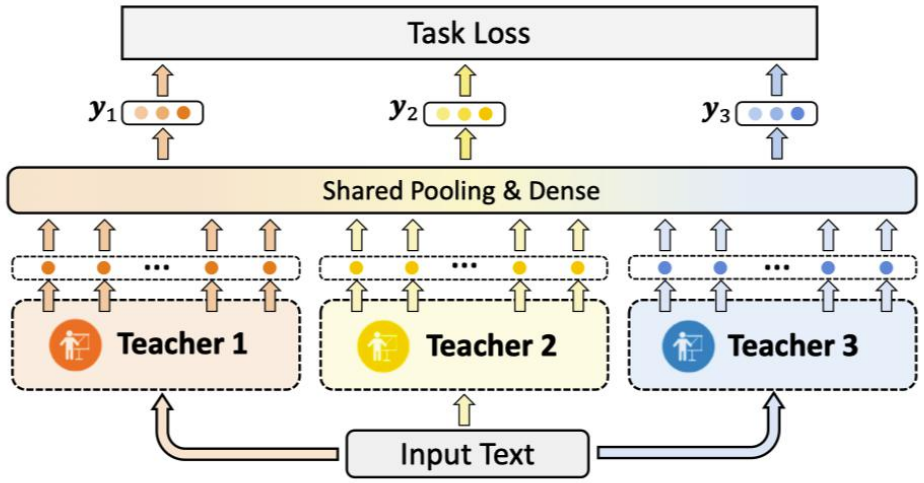


Figure 1: The multi-teacher co-finetuning framework

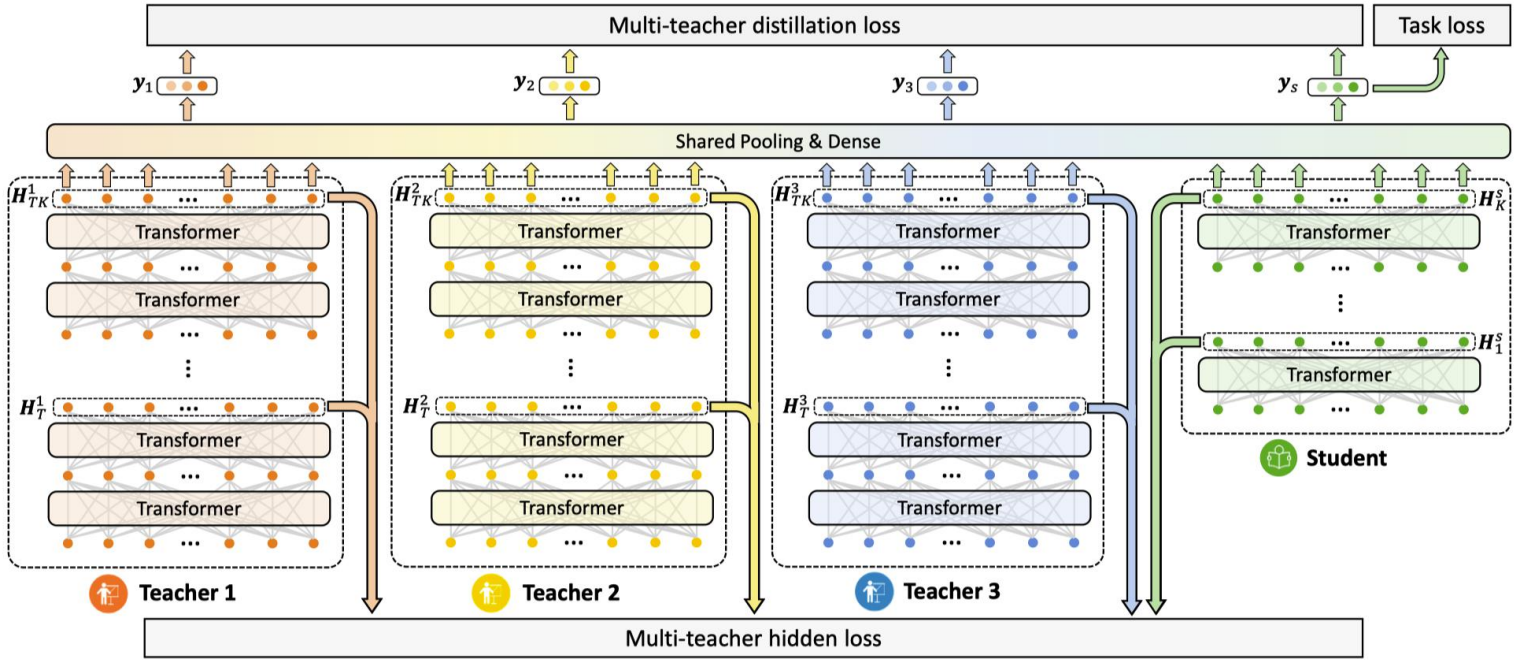
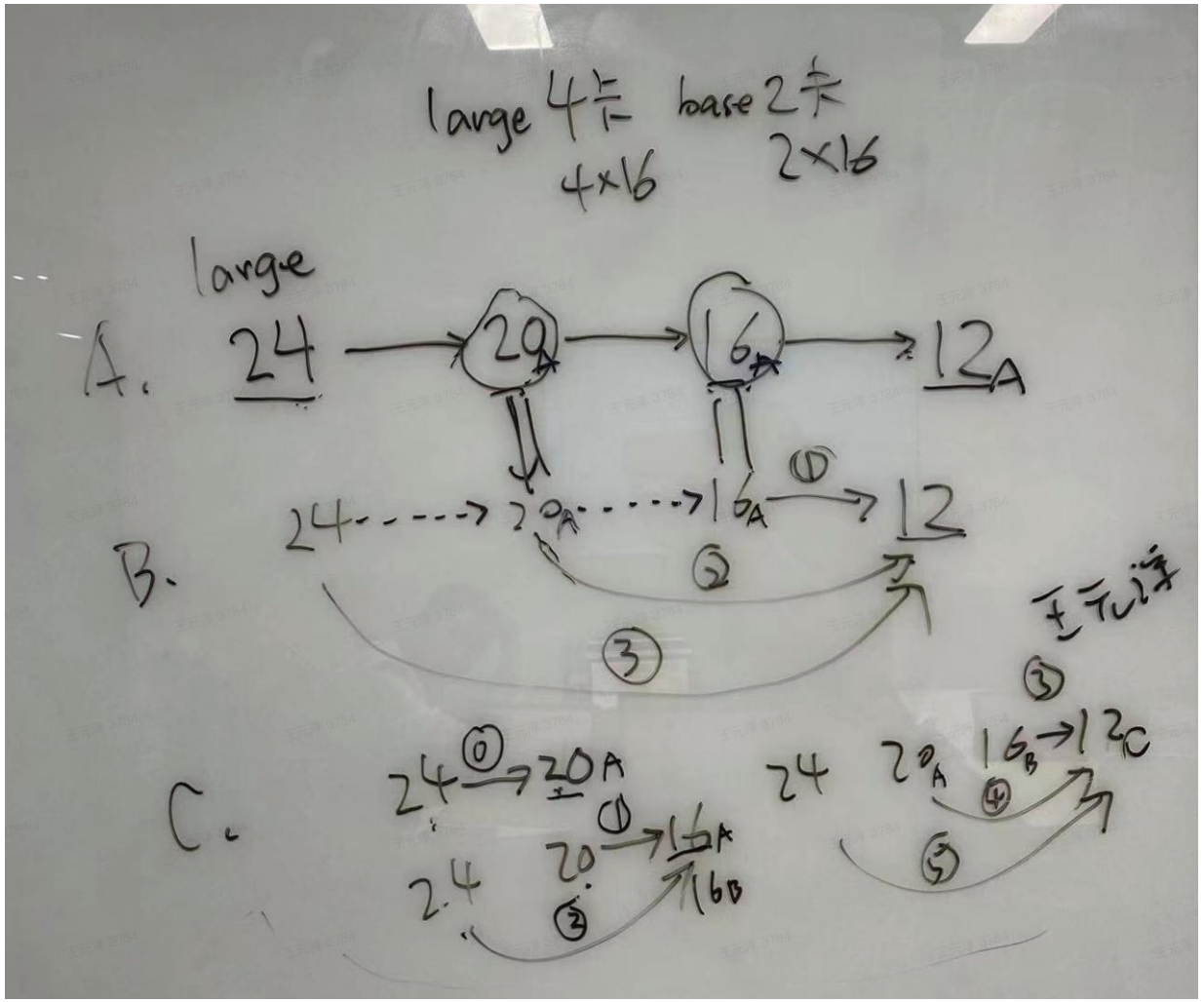


Figure 2: The multi-teacher knowledge distillation framework in MT-BERT.

3.3 多teacher

- 3.3.2 迭代方法



目录

4. 更多想法 & 挑战

- 超大模型超大gap：选定具体任务
  当前目标：GLM-10B -> 2B~7B
- data free
- 生成式LM的蒸馏

END

thx 4 listening