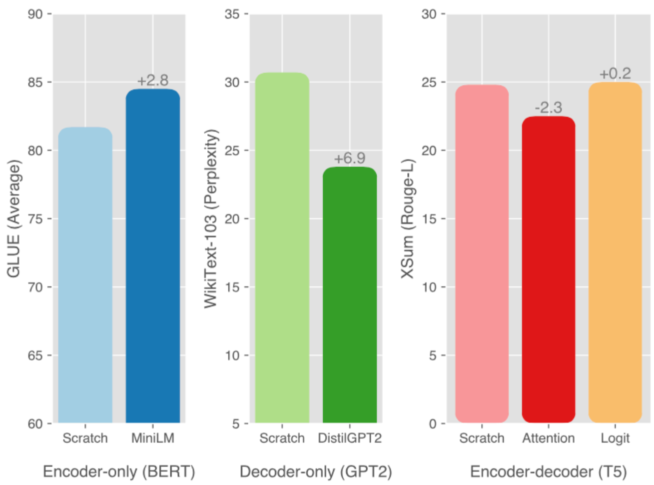
Task-agnostic Distillation for Encoder-decoder Language Models

Chen Zhang[1], Yang Yang[2], Qiuchi Li[3], Jingang Wang[2], Dawei Song[1*]
[1]Beijing Institute of Technology, [2]Meituan NLP, [3]University of Copenhagen *Corresponding author

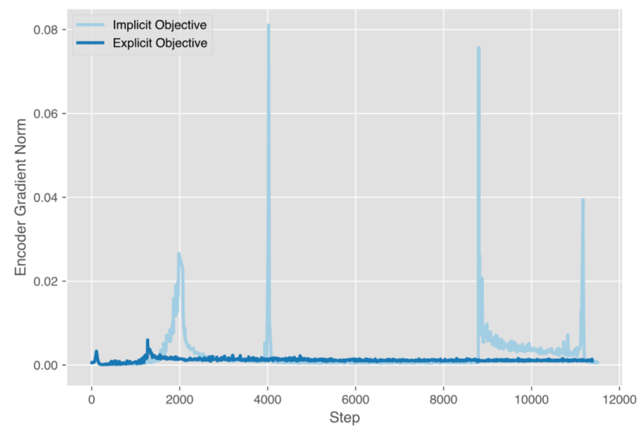
Motivation



- previous studies on task-agnostic distillation mainly focus on encoder-only language models (e.g. BERT) or decoder-only language models (e.g., GPT2).
- while attention-based distillation is suitable for BERT (e.g., MiniLM) and logit-based distillation is suitable for GPT2 (e.g., DistilGPT2), they fail to make distilled encoder-decoder language models surpass pretraining-from-scratch baseline.

Method

- encoder-decoder interplay is important, without which the distillation would be rather unstable:
  - *a gradient perspective by contrasting explicit objective involving the interplay to implicit objective, where explicit one admits stable training.*



$$\mathcal{L}^{\text{Imp}} = \mathcal{L}^{\text{Logit}}(\mathcal{S};\mathcal{T},\mathcal{D}_{\mathbf{Z}}) + \mathcal{L}^{\text{SelfAttn}}(\mathcal{S};\mathcal{T},\mathcal{D}_{\mathbf{Z}}),$$

$$\mathcal{L}^{\text{Exp}} = \mathcal{L}^{\text{Logit}}(\mathcal{S};\mathcal{T},\mathcal{D}_{\mathbf{Z}}) + \mathcal{L}^{\text{SelfAttn}}(\mathcal{S};\mathcal{T},\mathcal{D}_{\mathbf{Z}}) + \mathcal{L}^{\text{CrossAttn}}(\mathcal{S};\mathcal{T},\mathcal{D}_{\mathbf{Z}},\mathcal{D}_{\mathbf{E}})$$

$$\mathcal{L}(\mathcal{S};\mathcal{T},\mathcal{D}_{\mathbf{Z}},\mathcal{D}_{\mathbf{X}}) = \mathcal{L}^{\text{Logit}}(\mathcal{S};\mathcal{T},\mathcal{D}_{\mathbf{Z}}) + \mathcal{L}^{\text{SelfAttn}}(\mathcal{S};\mathcal{T},\mathcal{D}_{\mathbf{Z}}) + \mathcal{L}^{\text{EncSelfAttn}}(\mathcal{S};\mathcal{T},\mathcal{D}_{\mathbf{X}}),$$

- based on the observation, two explicit objectives are proposed:
  - *one considering decoder self-attention and decoder cross-attention (MiniEnD-D), and another considering decoder self-attention and encoder self-attention (MiniEnD-E).*

Results

Method	GFLOPs		SST-2 Acc	MRPC F1	STS-B SpCorr	QQP F1	MNLI-m/mm Acc	QNLI Acc	RTE Acc	GLUE Score
$T5_{base}$	25.4	1×	94.6	93.0	90.0	88.9	86.7/86.8	92.9	74.7	88.5
$T5_{6L;384H}$	3.18	3~8×	92.2	90.2	86.0	87.3	81.2/81.7	88.2	**70.0**	84.6
$MiniDisc_{5\%}$ ①	7.80		**93.8**	89.8	85.3	86.7	**82.9**/82.7	**89.2**	64.6	84.4
$MlmKD_{6L;384H}$	3.18		92.3	88.7	86.2	87.5	81.6/82.1	88.2	67.9	84.3
$MiniLM_{6L;384H}$	3.18		92.1	89.6	85.2	87.0	81.2/81.5	88.0	68.6	84.1
$MlmKD{+}MiniLM_{6L;384H}$	3.18		92.4	89.2	86.0	87.3	81.7/82.1	89.1	67.9	84.5
$MiniEnD\text{-}D_{6L;384H}$	3.18	8×	92.1	**90.6**	85.8	**87.7**	81.8/82.3	89.0	68.6	84.7
w/o $\mathcal{L}^{\text{Logit}}$	3.18		92.2	90.1	**86.6**	87.6	82.2/82.8	89.1	68.6	84.9
$MiniEnD\text{-}E_{6L;384H}$	3.18		92.7	90.0	86.1	87.4	81.8/82.1	88.8	69.3	84.8
w/o $\mathcal{L}^{\text{Logit}}$	3.18		92.3	89.9	**86.6**	**87.7**	82.5/**83.1**	**89.2**	69.0	**85.0**

① MiniDisc is distilled from $T5_{xlarge}$, and owns larger GFLOPs.

- comprehensive results demonstrate the improved effectiveness, see more results in our paper.

Conclusions

- in this paper, we aim to provide a path that suc- cessfully tackles the distillation of encoder-decoder LMs, which fails most previous methods in the area. We find through a pilot study that the encoder- decoder interplay is a key component that should be aligned in the distillation so that the distilled encoder-decoder LMs are promising. Based on the idea, we propose two directions that the encoder- decoder interplay alignment can be incorporated and verify their effectiveness on a language understanding benchmark and two abstractive summarization datasets.
- this work is funded in part by the Natural Science Foundation of China (grant no: 62376027) and Beijing Municipal Natural Science Foundation (grant no: 4222036 and IS23061).