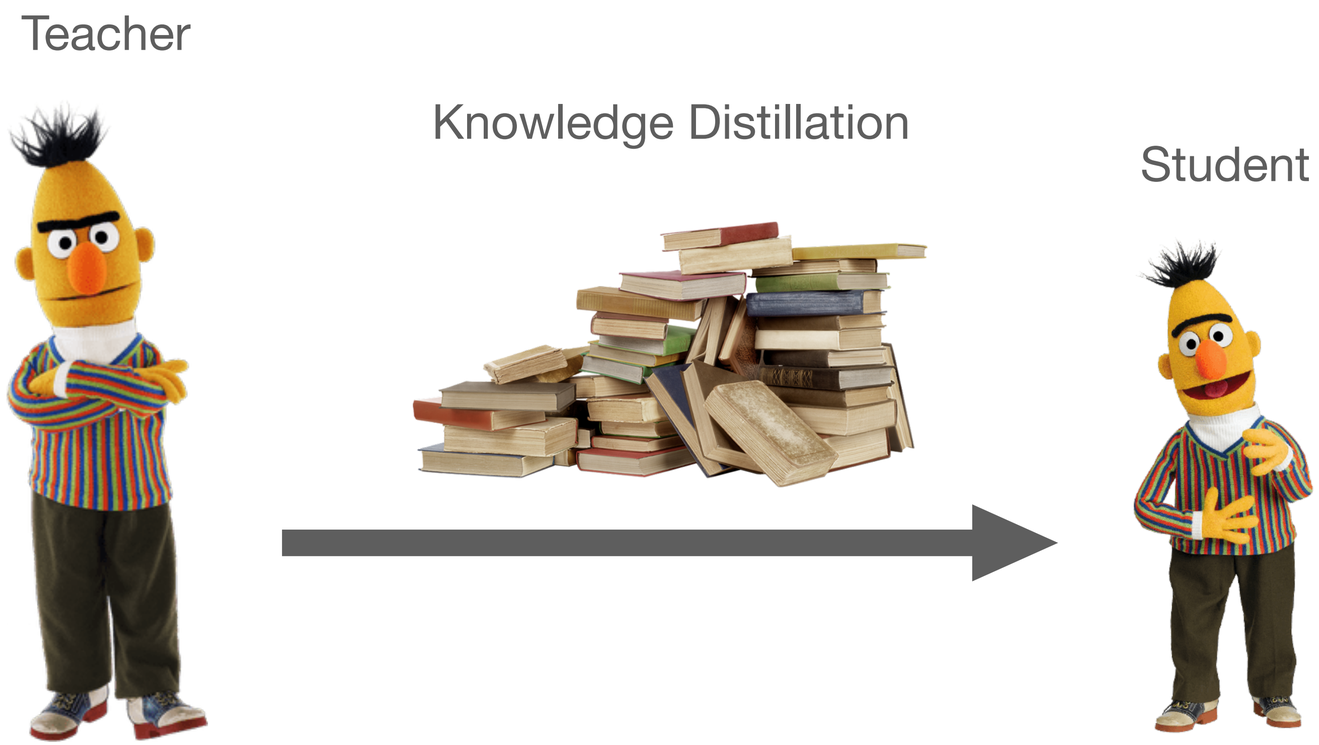
Task-agnostic Distillation of Encoder-decoder Language Models

Chen Zhang, Yang Yang, Quichi Li, Jingang Wang, Dawei Song

Language Model Distillation

Teacher-student Paradigm

- Language model (LM) distillation aims at reducing inference compute by distilling the large LM into a small LM under a teacher-student paradigm.



Language Model Distillation

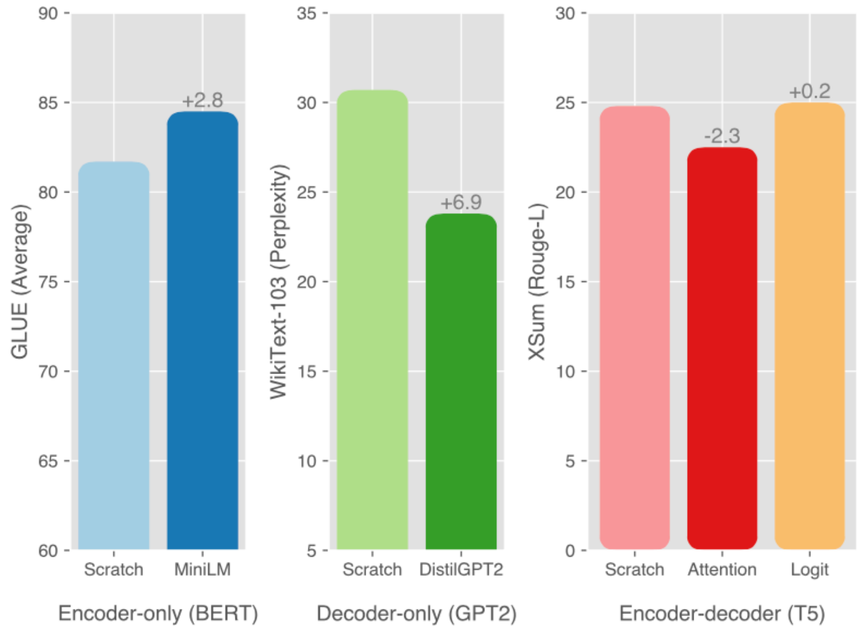
Existing Methods

- Task-specific distillation with finetuning data (e.g., MRPC).
  - KD (Hinton, et al.)
  - MiniDisc (Zhang, et al.)
- Task-agnostic distillation with pretraining data (e.g., Wikipedia).
  - MiniLM (Wang, et al.): attention distribution-based
  - DistilGPT2 (Sanh et al.): logit distribution-based
- Task-agnostic distillation is commonly viewed as a better choice.

Failure of Distillation

Encoder-decoder Language Models

- In distilling a base-scale teacher to a 6-layer student, we have found that previous studies that are applicable either to encoder-only language models (e.g., BERT) or to decoder-only language models (e.g., GPT2) fail to handle encoder-decoder language models (e.g., T5).



Encoder-decoder Interplay

Gradient Perspective

- We suspect previous studies lack a component accounting for encoder-decoder interplay with only (MiniLM-like) logit distribution-based or (DistilGPT2-like) attention distribution-based distillation. Distinguishing from previous implicit objectives, we propose explicit objectives that involve encoder-decoder interplay.

$$\mathcal{L}^{\text{SelfAttn}}(\mathcal{S};\mathcal{T},\mathcal{D}_{\mathbf{Z}}) = \mathbb{E}_{\mathbf{Z}\sim\mathcal{D}_{\mathbf{Z}}} \sum_{k=1}^{R} \mathsf{KL}(\mathsf{Reln}(\mathbf{Z};{}^{\mathcal{T}}\mathbf{W}_k^{\mathsf{Q}}), \mathsf{Reln}(\mathbf{Z};{}^{\mathcal{S}}\mathbf{W}_k^{\mathsf{Q}})) + \mathsf{KL}(\mathsf{Reln}(\mathbf{Z};{}^{\mathcal{T}}\mathbf{W}_k^{\mathsf{K}}), \mathsf{Reln}(\mathbf{Z};{}^{\mathcal{S}}\mathbf{W}_k^{\mathsf{K}})) + \mathsf{KL}(\mathsf{Reln}(\mathbf{Z};{}^{\mathcal{T}}\mathbf{W}_k^{\mathsf{V}}), \mathsf{Reln}(\mathbf{Z};{}^{\mathcal{S}}\mathbf{W}_k^{\mathsf{V}})),$$

$$\mathsf{Reln}(\mathbf{Z};{}^{\mathcal{T}}\mathbf{W}_k^{\mathsf{Q}}) = \mathsf{Softmax}(\mathbf{Z}{}^{\mathcal{T}}\mathbf{W}_k^{\mathsf{Q}}{}^{\mathcal{T}}\mathbf{W}_k^{\mathsf{Q}\top}\mathbf{Z}^{\top}/d^{\mathsf{R}}),$$

$$\mathcal{L}^{\text{CrossAttn}}(\mathcal{S};\mathcal{T},\mathcal{D}_{\mathbf{Z}},\mathcal{D}_{\mathbf{E}}) = \mathbb{E}_{\mathbf{Z}\sim\mathcal{D}_{\mathbf{Z}},\mathbf{E}\sim\mathcal{D}_{\mathbf{E}}} \sum_{k=1}^{R} \mathsf{KL}(\mathsf{Reln}(\mathbf{Z};{}^{\mathcal{T}}\mathbf{W}_k^{\mathsf{Q}'}), \mathsf{Reln}(\mathbf{Z};{}^{\mathcal{S}}\mathbf{W}_k^{\mathsf{Q}'})) + \mathsf{KL}(\mathsf{Reln}(\mathbf{E};{}^{\mathcal{T}}\mathbf{W}_k^{\mathsf{K}'}), \mathsf{Reln}(\mathbf{E};{}^{\mathcal{S}}\mathbf{W}_k^{\mathsf{K}'})) + \mathsf{KL}(\mathsf{Reln}(\mathbf{E};{}^{\mathcal{T}}\mathbf{W}_k^{\mathsf{V}'}), \mathsf{Reln}(\mathbf{E};{}^{\mathcal{S}}\mathbf{W}_k^{\mathsf{V}'})),$$

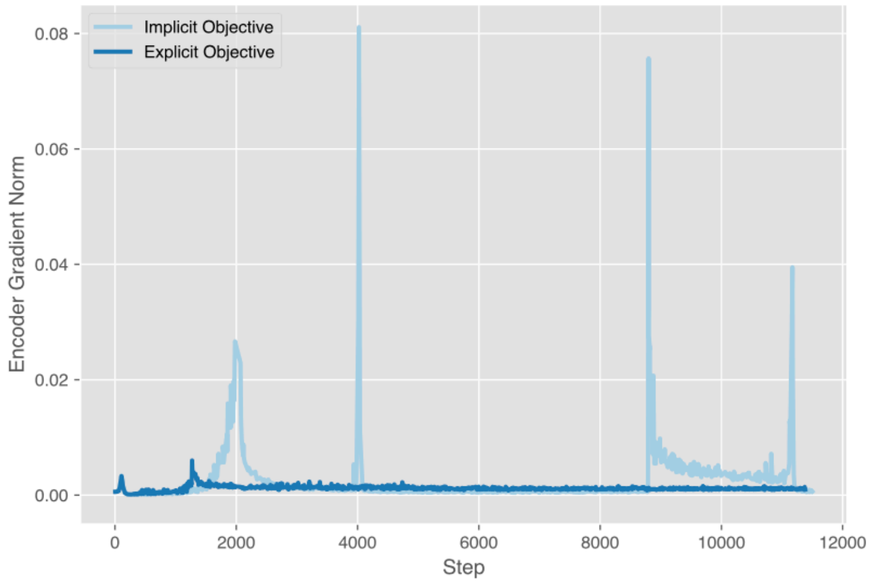
$$\mathcal{L}^{\text{Imp}} = \mathcal{L}^{\text{Logit}}(\mathcal{S};\mathcal{T},\mathcal{D}_{\mathbf{Z}}) + \mathcal{L}^{\text{SelfAttn}}(\mathcal{S};\mathcal{T},\mathcal{D}_{\mathbf{Z}}),$$

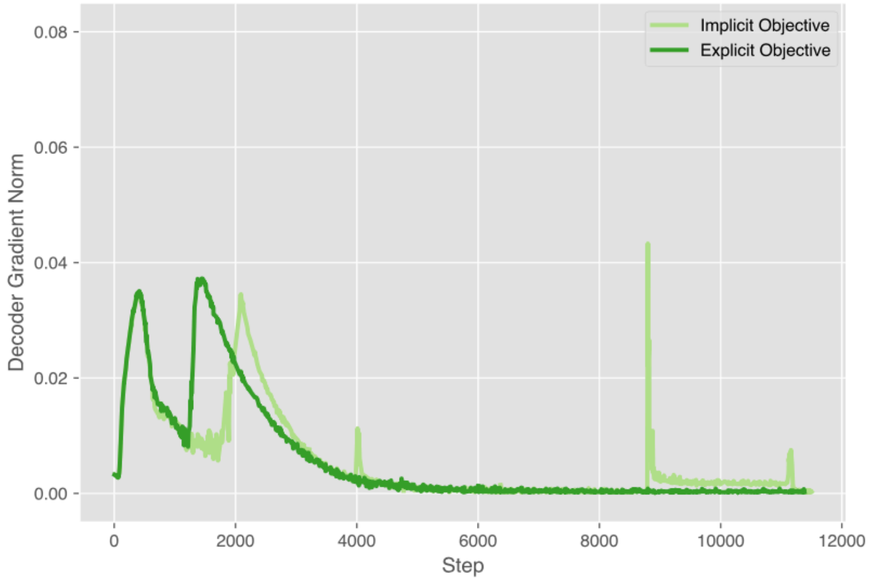
$$\mathcal{L}^{\text{Exp}} = \mathcal{L}^{\text{Logit}}(\mathcal{S};\mathcal{T},\mathcal{D}_{\mathbf{Z}}) + \mathcal{L}^{\text{SelfAttn}}(\mathcal{S};\mathcal{T},\mathcal{D}_{\mathbf{Z}}) + \mathcal{L}^{\text{CrossAttn}}(\mathcal{S};\mathcal{T},\mathcal{D}_{\mathbf{Z}},\mathcal{D}_{\mathbf{E}})$$

Encoder-decoder Interplay

Gradient Perspective

- We have uncovered than implicit objectives would impose unstable gradient norms, thereby unstable training. In contrast, explicit objectives are much more stable.





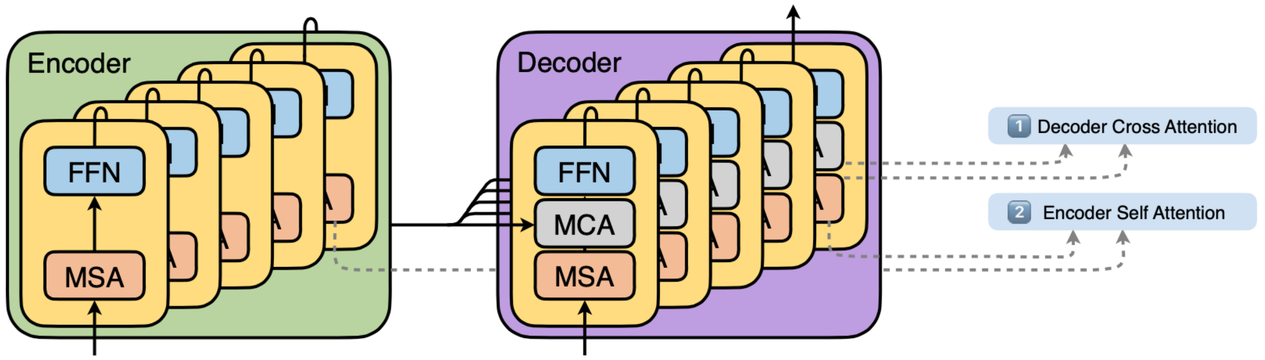
Encoder-Decoder Interplay

MiniEnD

- Based on the observation, we put forward two implementations of the explicit objectives. One is based directly on encoder-decoder cross-attention, and the other is based on both encoder and decoder self-attention.

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

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



Experiments

Setup

- Distillation on C4 for T5, on OpenWebText for BART.
- Finetuning on GLUE (sequence and sequence-pair classification), CNN/DailyMail and XSum (summarization).
- T5-based and BART-base as teachers.

Experiments

GLUE

- The distillation cannot surpass pretraining-from-scratch baseline on GLUE until MiniEnD.

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-$D_{6L;384H}$	3.18	8×	92.1	**90.6**	85.8	**87.7**	81.8/82.3	89.0	68.6	84.7
w/o $\mathcal{L}^{Logit}$	3.18		92.2	90.1	**86.6**	87.6	82.2/82.8	89.1	68.6	84.9
MiniEnD-$E_{6L;384H}$	3.18		92.7	90.0	86.1	87.4	81.8/82.1	88.8	69.3	84.8
w/o $\mathcal{L}^{Logit}$	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.

Experiments

CNN/DailyMail & XSum

- Similar results are observed on CNN/DailyMail and XSum.

Method	GFLOPs		CNN/DailyMail Rg-1	CNN/DailyMail Rg-2	CNN/DailyMail Rg-L	XSum Rg-1	XSum Rg-2	XSum Rg-L
$T5_{base}$	25.4	1×	40.1	19.4	31.5	34.7	12.4	29.7
$T5_{6L;384H}$	3.18	8×	35.7	16.8	28.4	28.6	8.9	24.8
$MlmKD_{6L;384H}$	3.18	8×	36.0	17.0	28.7	28.9	**9.2**	25.0
$MiniLM_{6L;384H}$	3.18	8×	35.0	16.5	28.0	25.9	7.5	22.5
$MlmKD+MiniLM_{6L;384H}$	3.18	8×	35.8	17.0	28.7	29.0	9.1	25.1
MiniEnD-$D_{6L;384H}$	3.18	8×	**36.2**	17.2	**28.9**	**29.5**	**9.2**	**25.4**
w/o $\mathcal{L}^{Logit}$	3.18	8×	35.7	17.0	28.6	27.3	8.2	23.7
MiniEnD-$E_{6L;384H}$	3.18	8×	36.1	**17.3**	**28.9**	28.9	9.1	24.9
w/o $\mathcal{L}^{Logit}$	3.18	8×	35.8	17.1	28.7	27.2	8.0	23.6
$BART_{base}$	12.7	1×	39.4	18.5	30.6	36.9	14.7	31.9
$LogitKD_{3/1L;768H}$ ①	4.23	1~3×	38.0	16.0	25.2	32.9	12.4	26.9
$DQ\text{-}BART_{8bit}$ ②	12.7	1~3×	**42.4**	**19.3**	28.8	**38.2**	**15.7**	**30.7**
MiniEnD-$D_{6L;384H}$	3.18	4×	38.5	18.5	**29.7**	33.6	12.9	29.2

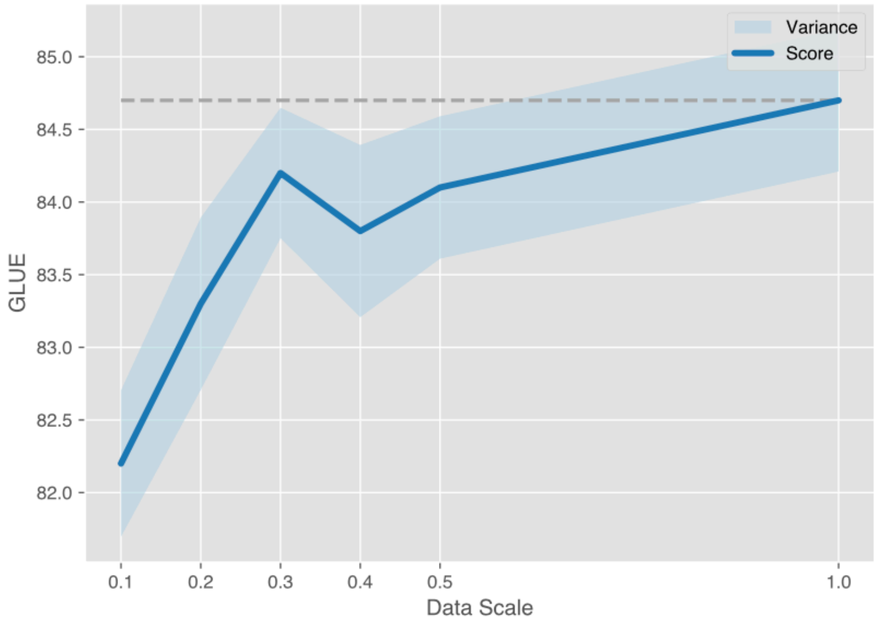
① LogitKD is distilled with an asymmetric layer setting, i.e., more encoder layers than decoder layers, for saved performance decline.

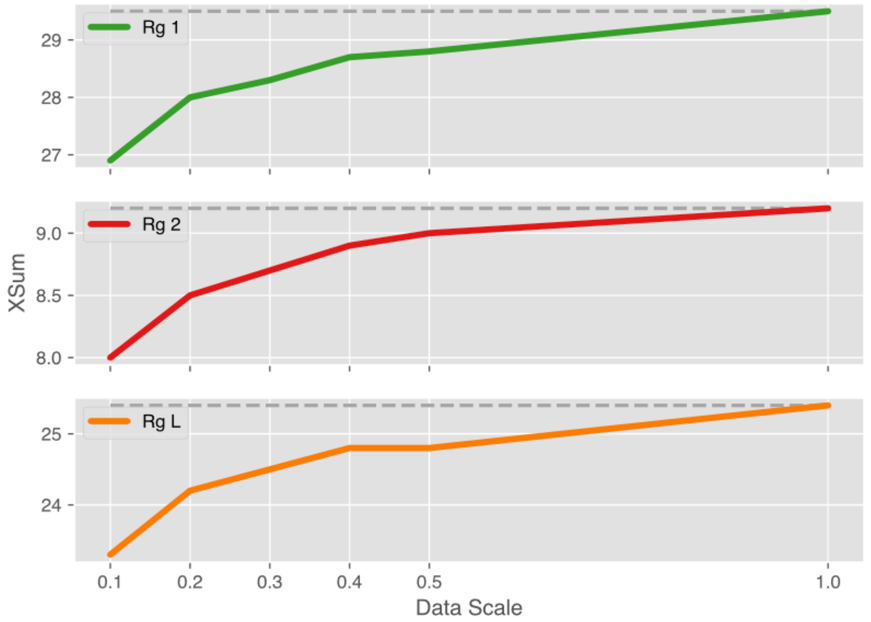
② DQ-BART only quantizes parameter precision to lower one, i.e., 8 bit, but does not reduce parameter amount. Quantization would not give any speedup in GFLOPs though nice reduction in model size.

Experiments

Data Scaling

- MiniEnD is also applicable when half data is used.





Experiments

Model Scaling

- Special methods should be attached to MiniEnd (e.g., TA) when teacher is scaled up.

Table 5: The results of model scaling using MiniEnD-D. $\Rightarrow$ denotes a distillation step, which should be operated sequentially otherwise $\{\}$ is prioritized. $\cdots \Rightarrow \cdots \Rightarrow \cdots$ indicates teacher assistant-based distillation and $\cdots \Rightarrow \{\cdots \Rightarrow \cdots\}$ indicates progressive distillation.

Method	GLUE Score	CNN/DailyMail Rg-1	CNN/DailyMail Rg-2	CNN/DailyMail Rg-L	XSum Rg-1	XSum Rg-2	XSum Rg-L
$T5_{6L;384H}$	84.6	35.7	16.8	28.4	28.6	8.9	24.8
$T5_{12L;384H}$	85.0	37.2	17.9	29.6	31.2	10.5	27.0
$T5_{base}$	88.5	40.1	19.4	31.5	34.7	12.4	29.7
$T5_{large}$	90.7	40.6	19.4	31.7	38.2	15.1	32.9
$T5_{xlarge}$	92.0	40.8	19.7	32.1	41.1	17.6	35.5
$T5_{base} \Rightarrow T5_{6L;384H}$	84.7	36.2	17.2	28.9	29.5	9.2	25.4
$T5_{large} \Rightarrow T5_{6L;384H}$	84.5	36.4	17.4	29.0	29.4	9.3	25.3
$T5_{xlarge} \Rightarrow T5_{6L;384H}$	84.2	36.1	17.2	28.8	29.1	9.1	25.1
$T5_{xlarge} \Rightarrow T5_{12L;384H} \Rightarrow T5_{6L;384H}$	84.6	36.6	17.5	29.2	29.2	9.1	25.1
$T5_{large} \Rightarrow T5_{12L;384H}$	85.5	38.3	18.4	30.4	32.4	11.2	27.9
$T5_{xlarge} \Rightarrow T5_{12L;384H}$	85.2	38.0	18.4	30.3	32.2	11.1	27.7
$T5_{xlarge} \Rightarrow \{T5_{large} \Rightarrow T5_{12L;384H}\}$	85.8	38.4	18.5	30.6	32.9	11.5	28.3

Conclusion

- 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.
- We further scale the distillation of encoder-decoder LMs to a 3B teacher that requires additional distillation steps. In this sense, we recommend future research to devote more efforts to exploring how large language models can be distilled.
- arXiv: https://arxiv.org/abs/2305.12330
- GitHub: https://github.com/GeneZC/MiniEnD
- Slides: https://genezc.github.io/assets/files/COLING2024_MiniEnD.pdf
- Thank you all!