AutoDistil: Neural Architecture Search for Distilling Large Language Models

https://aka.ms/autodistil

Dongkuan (DK) Xu, Subhabrata Mukherjee, Xiaodong Liu, Debadeepta Dey,

Wenhui Wang, Xiang Zhang, Ahmed Hassan Awadallah, Jianfeng Gao

NC STATE UNIVERSITY

Microsoft Research

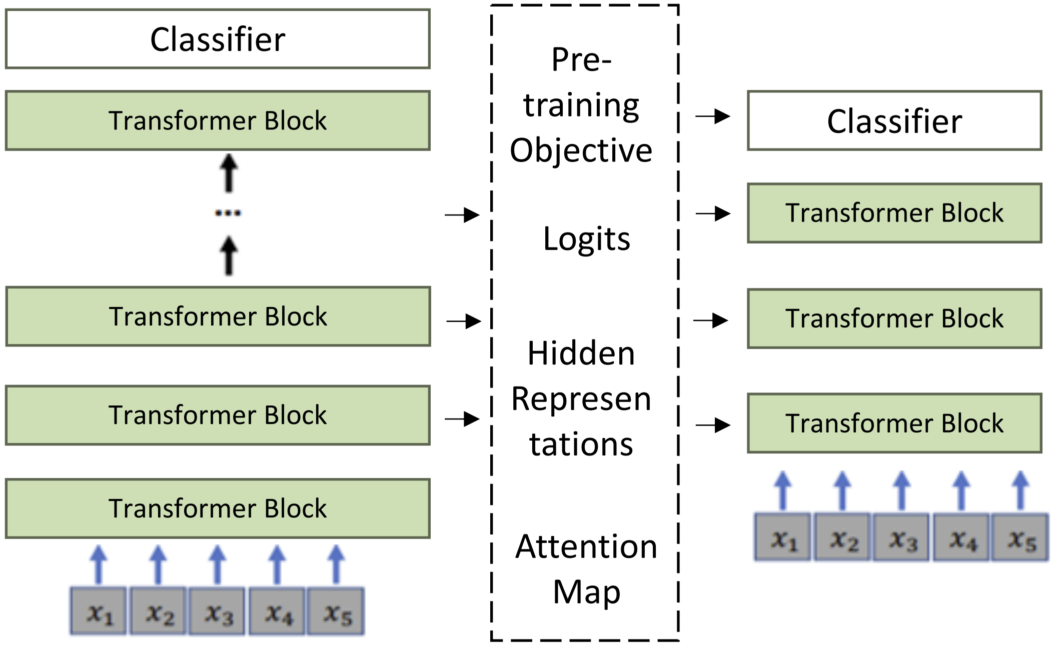
PennState

Knowledge Distillation of BERT

Use the large over-parameterized model to distil a small model

Multi-Objective Knowledge Distillation:

- Teacher Logits
- Multi-layer hidden state transfer
- Attention Map Transfer

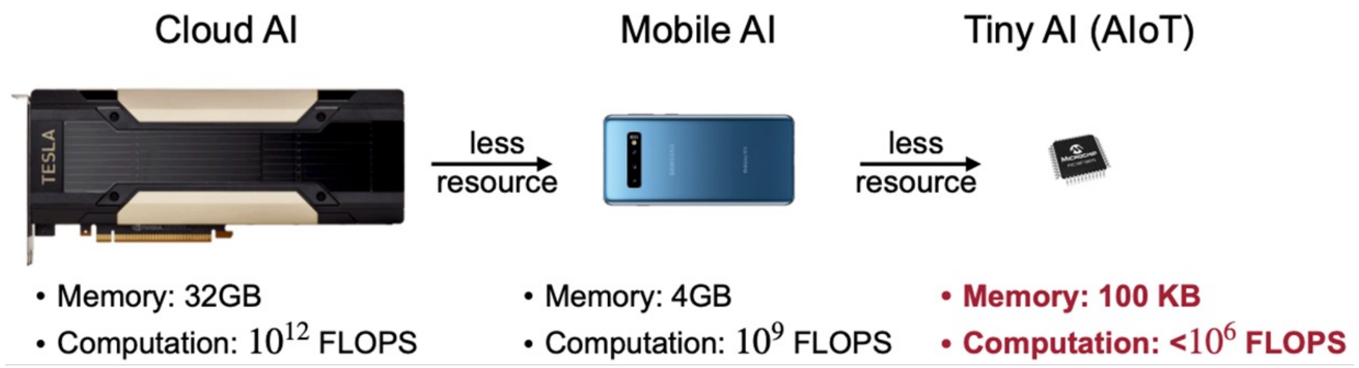


XtremeDistil: Multi-stage Distillation. ACL 2020. Mukherjee and Awadallah https://aka.ms/xtremedistil
TransferDistil: Task Transfer for Task-agnostic Distillation. Mukherjee et al., 2021
MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. NeurIPS 2020. Wang et al

What are the challenges?

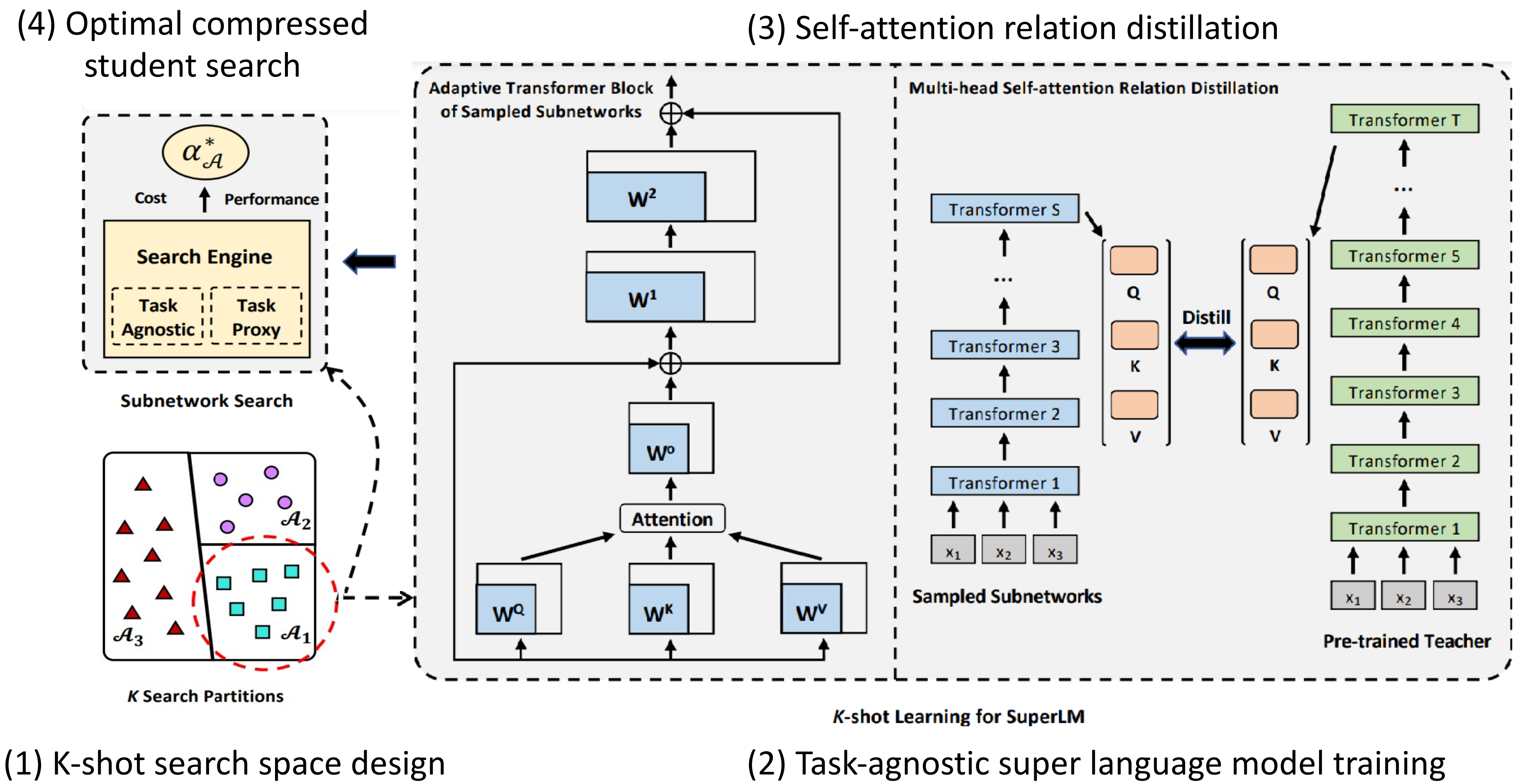
- Small model architectures are hand-designed
  - Requires several trials
  - Relies on pre-specified compression rates
  - Re-running distillation with computational budget change

- More, one size does not fit all



Source: https://ofa.mit.edu/

AutoDistil with Neural Architecture Search



https://aka.ms/AutoDistil

Search Space Design

- Searchable transformer components
- Inductive bias
- Search space partition

Three sub-spaces

Components

	SuperLM$_{\text{Tiny}}$	SuperLM$_{\text{Small}}$	SuperLM$_{\text{Base}}$	BERT
#Subnets	256	256	256	N/A
#Layers	(4, 7, 1)	(9, 12, 1)	(9, 12, 1)	12
#Hid_dim	(128, 224, 32)	(256, 352, 32)	(544, 640, 32)	768
MLP Ratio	(2.0, 3.5, 0.5)	(2.5, 4.0, 0.5)	(2.5, 4.0, 0.5)	4.0
#Heads	(7, 10, 1)	(7, 10, 1)	(9, 12, 1)	12
#FLOPs	40-367M	0.5-2.1G	2.1-7.9G	11.2G
#Params	4-10M	12-28M	39-79M	109M

Cover different ranges of #FLOPs & #Para

Each tuple represents the lowest value, highest value, and steps for component

AutoDistil vs. Manually Designed Distilled Models

Model (Metric)	#FLOPs (G)	#Para (M)	MNLI-m (Acc)	QNLI (Acc)	QQP (Acc)	SST-2 (Acc)	CoLA (Mcc)	MRPC (Acc)	RTE (Acc)	Average
$BERT_{BASE}$ Devlin et al. [2019] (teacher)	11.2	109	84.5	91.7	91.3	93.2	58.9	87.3	68.6	82.2
$BERT_{SMALL}$ Turc et al. [2019]	5.66	66.5	81.8	89.8	90.6	91.2	53.5	84.9	67.9	80.0
Truncated BERT Williams et al. [2018]	5.66	66.5	81.2	87.9	90.4	90.8	41.4	82.7	65.5	77.1
DistilBERTSanh et al. [2019]	5.66	66.5	82.2	89.2	88.5	91.3	51.3	87.5	59.9	78.6
TinyBERT Jiao et al. [2020]	5.66	66.5	83.5	90.5	90.6	91.6	42.8	88.4	72.2	79.9
MINILM Williams et al. [2018]	5.66	66.5	84.0	91.0	91.0	92.0	49.2	88.4	71.5	81.0
$\texttt{AutoDistil}_{\text{Agnostic}}$	2.13	26.8	82.8	89.9	90.8	90.6	47.1	87.3	69.0	79.6
$\texttt{AutoDistil}_{\text{Proxy}_B}$	4.40	50.1	83.8	90.8	91.1	91.1	55.0	88.8	71.9	81.7
$\texttt{AutoDistil}_{\text{Proxy}_S}$	2.02	26.1	83.2	90.0	90.6	90.1	48.3	88.3	69.4	79.9
$\texttt{AutoDistil}_{\text{Proxy}_T}$	0.27	6.88	79.0	86.4	89.1	85.9	24.8	78.5	64.3	72.6

Model	#Layers	#Hid	Ratio	#Heads	#FLOPs	#Para
$BERT_{BASE}$	12	768	4	12	11.2G	109M
MINILM	6	768	4	6	5.66G	66.5M
$\texttt{AutoDis.}_{\text{Agnostic}}$	11	352	4	10	2.13G	26.8M
$\texttt{AutoDis.}_{\text{Proxy}_B}$	12	544	3	9	4.40G	50.1M
$\texttt{AutoDis.}_{\text{Proxy}_S}$	11	352	4	8	2.02G	26.1M
$\texttt{AutoDis.}_{\text{Proxy}_T}$	7	160	3.5	10	0.27G	6.88M

AutoDistil Optimal Architectures

AutoDistil Variable Compression

- AutoDistil generates multiple students with variable computational cost
- Given any SOTA compressed model, AutoDistil finds students with better trade-off (FLOPs vs. Accuracy)