Less is More: Domain Adaptation with Lottery Ticket for Reading Comprehension

Haichao Zhu, Zekun Wang, Heng Zhang, Ming Liu, Sendong Zhao, Bing Qin

Harbin Institute of Technology, Harbin, China

Reading Comprehension Task

Context: *"Victorian schools are either publicly or privately funded. Public schools, also known as state or government schools, are funded and run directly by the Victoria Department of Education. Students do not pay tuition fees, but some extra costs are levied. Private fee-paying schools include parish schools similar to British public schools …"*

Question: *What organization runs the public schools in Victoria?* **Answer:** *Victoria Department of Education*

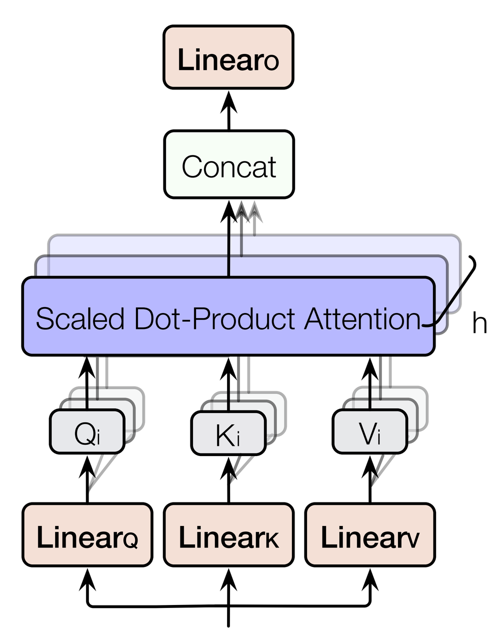
Motivation

- Large-scale annotation data is **expensive** to acquire, but **required** to achieve superior performance 📈 on reading comprehension task. **Few-shot** domain adaptation is a **practical** workaround.
- Pre-trained models with **millions of parameters** are **over-parameterized** and prone to easily **overfit tiny-scale data** which hinders generalization.
- **Not all parameters are equally ⚖️ important!**

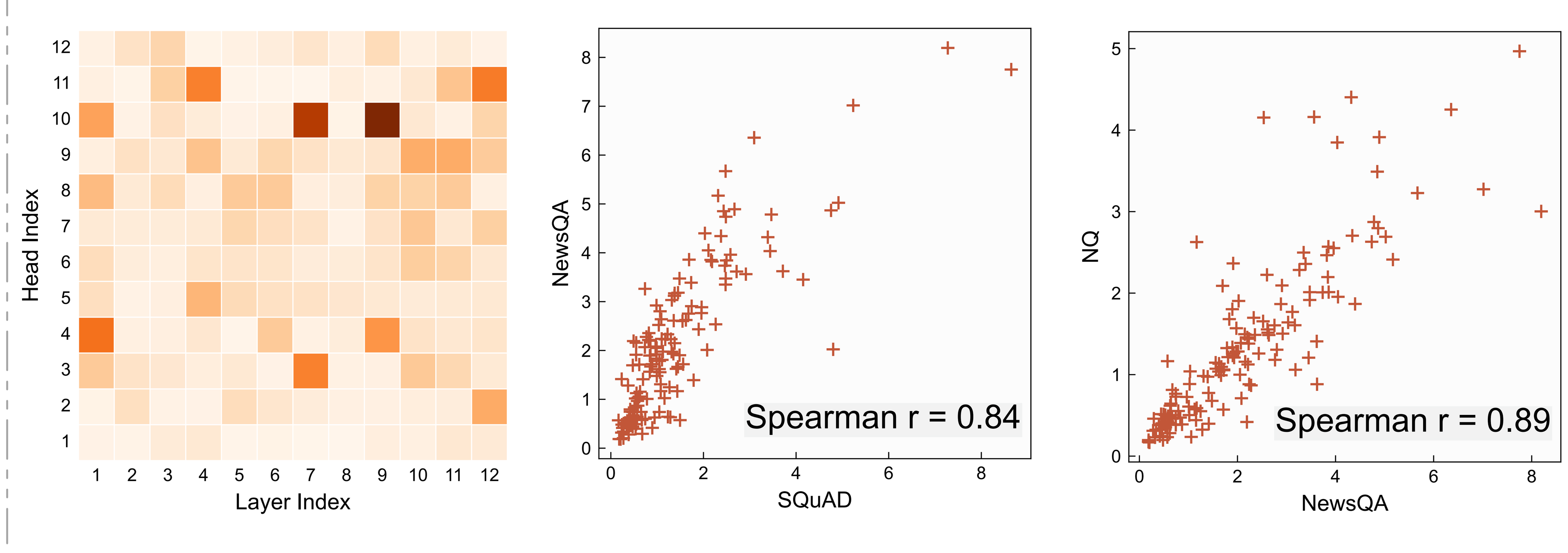
Background

Lottery Ticket Hypothesis: *Existing small and sparse subnetworks that rival the original network in performance, when trained in isolation from "lucky" initializations.*

Transformer

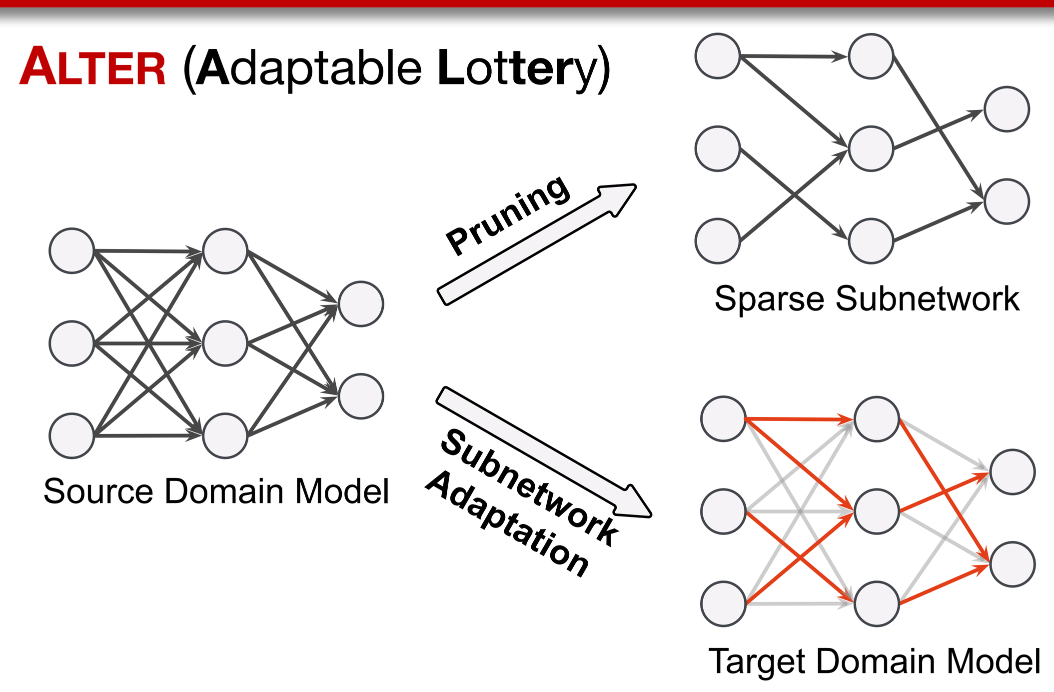


Self-Attention Head Analysis



Method

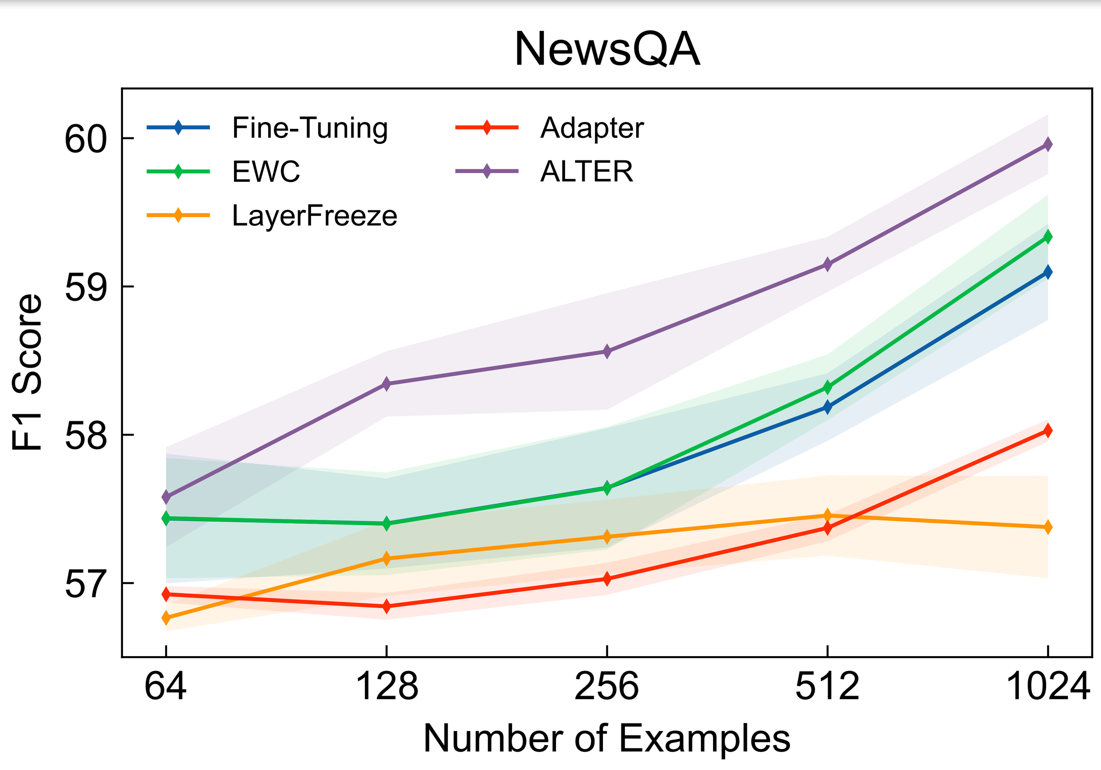
ALTER (**A**daptable **L**ot**ter**y)

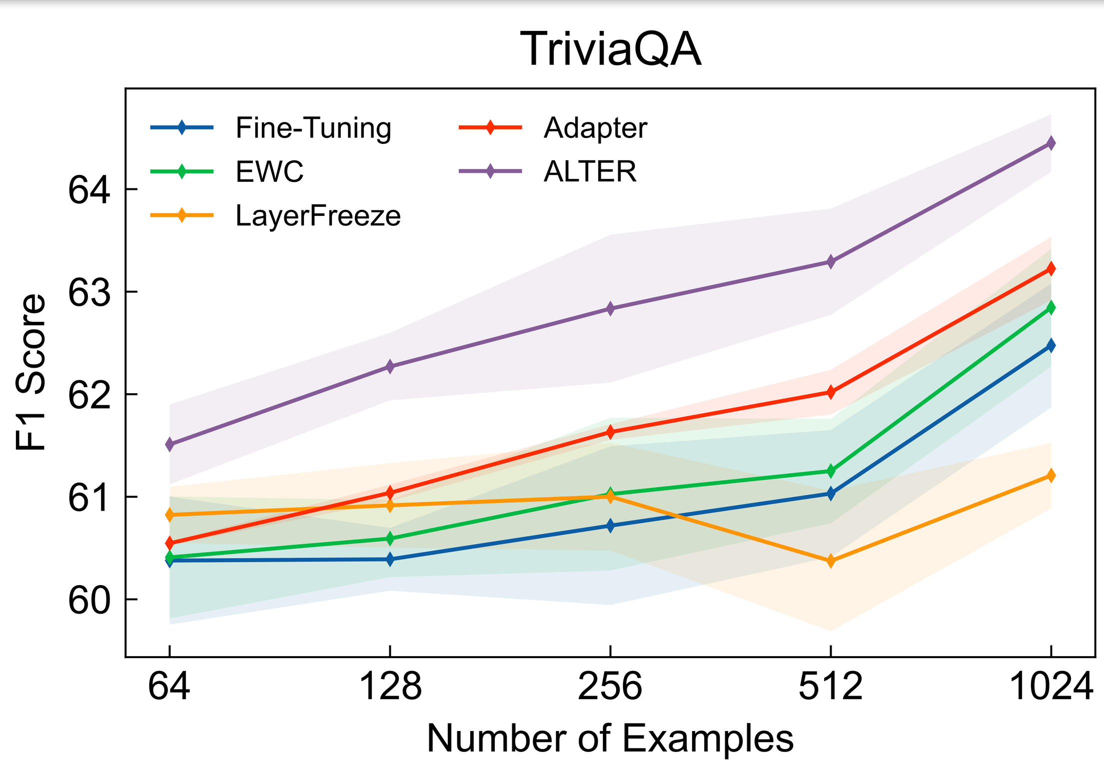


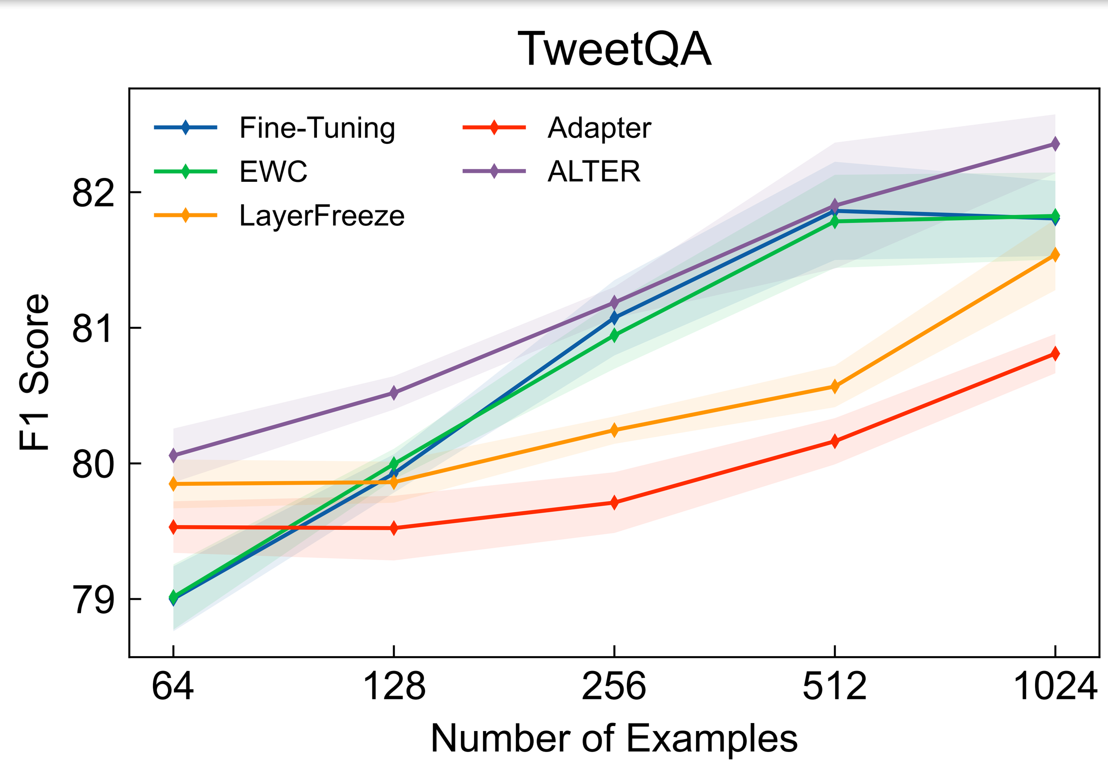
Algorithm 1 Identifying the Lottery Subnetwork with Self-Attention Head Importance

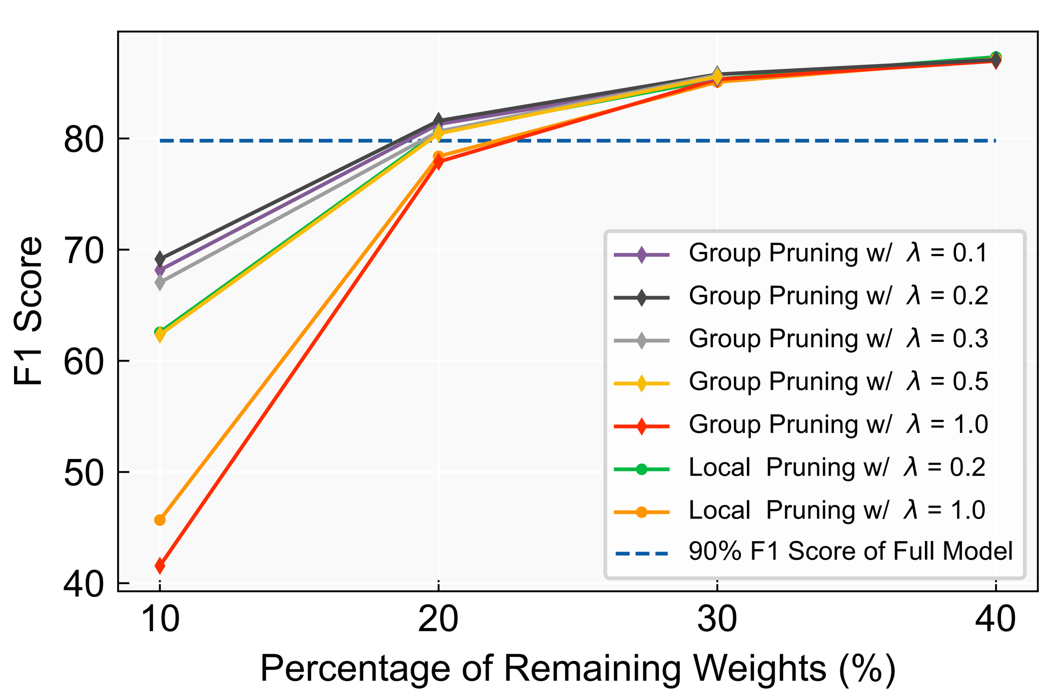
Require:
1: Source domain model $\mathcal{F}(\mathrm{x}; \mathbf{M} \odot \theta_0)$
2: Initial pruning mask $\mathbf{M} = 1^{|\theta_0|}$
3: Target sparsity s, pruning frequency ∇t and steps N
4: Importance factor λ
5: **for** $n \leftarrow 1$ **to** N **do**
6: Estimate attention head importance I_n
7: $\hat{I}_n \leftarrow \lambda + (1-\lambda)\frac{I_n - \min(I_n)}{\max(I_n)-\min(I_n)}$ ▷ normalize
8: Trim magnitudes with normalized importance score, $\hat{\theta}_{(n-1)\nabla t} \leftarrow AttrMagnitude(\theta_{(n-1)\nabla t}, \hat{I}_n)$
9: $s_n \leftarrow s - s(1-\frac{n}{N})^2$ ▷ sparsity of step n
10: Prune the lowest magnitudes parameters in group from $\hat{\theta}_{(n-1)\nabla t}$ to sparsity s_n
11: Update the pruning mask $\mathbf{M}$
12: Train the model for ∇t steps, producing $\mathcal{F}(\mathrm{x}; \mathbf{M} \odot \theta_{n\nabla t})$
13: **end for**
14: Train the model util stopping criterion is met, producing $\mathcal{F}(\mathrm{x}; \mathbf{M} \odot \theta_T)$
15: **return** Lottery Subnetwork $\mathbf{M}$

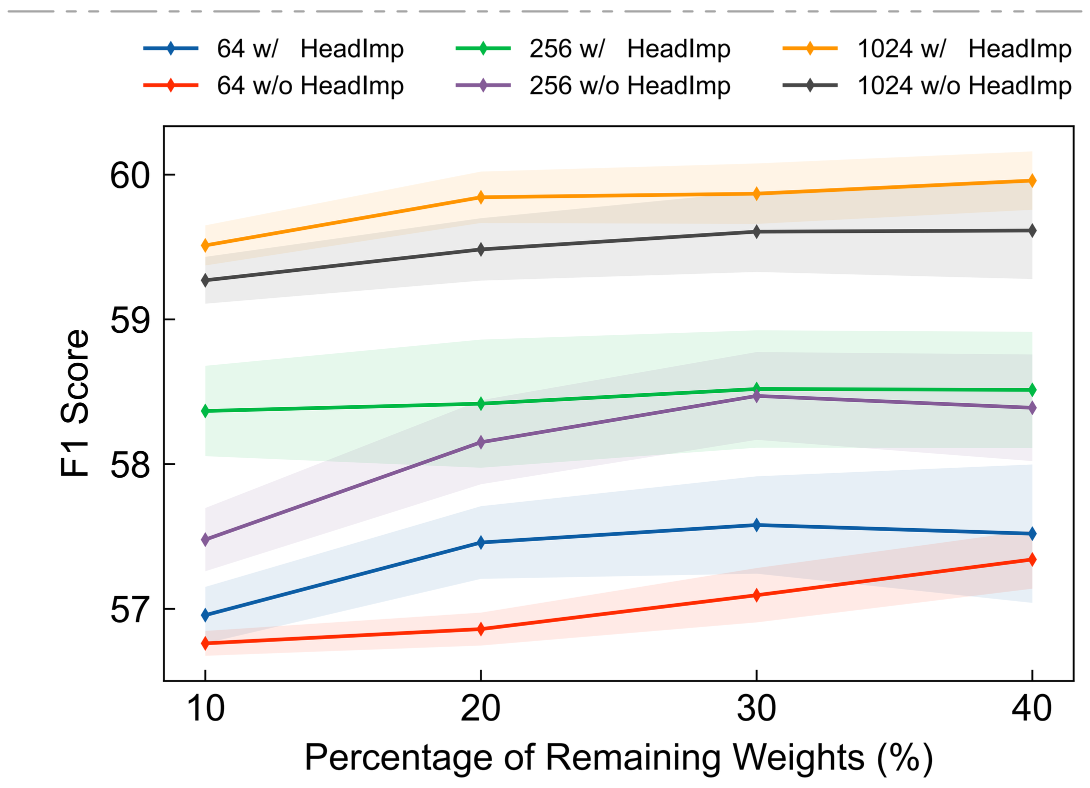
Results & Analysis











Method	NewsQA EM/F1	TriviaQA EM/F1	TweetQA EM/F1
Fine-Tuning	40.59/57.40	53.13/60.39	68.23/79.93
Random	40.98/57.72	54.45/61.98	68.57/80.24
Magnitude	40.76/57.56	54.10/62.11	68.76/80.21
Salvage	40.86/57.67	54.39/61.75	68.82/80.24
AttrHead	41.31/58.08	54.35/61.80	68.88/**80.39**
Alter	**41.38/58.11**	**54.60/62.21**	**68.89**/80.35

Table 1: Performance of different subnetwork identification methods on three target domain datasets with 128 examples. The number of parameters are 21M, which corresponds to 25% of the full model size. **Structured attention head importance scores help identify better lottery subnetworks.**