Doge Tickets: Uncovering Domain-general Language Models by Playing Lottery Tickets

Yi Yang#, Chen Zhang#, Benyou Wang, Dawei Song

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

香港中文大學
The Chinese University of Hong Kong

Outline

- Background: Out-of-domain Generalization
- Motivation
- Method: Identifying the Doge Tickets
- Experiments
- Results & Analysis
- Conclusion

Background

- Out-of-domain Generalization

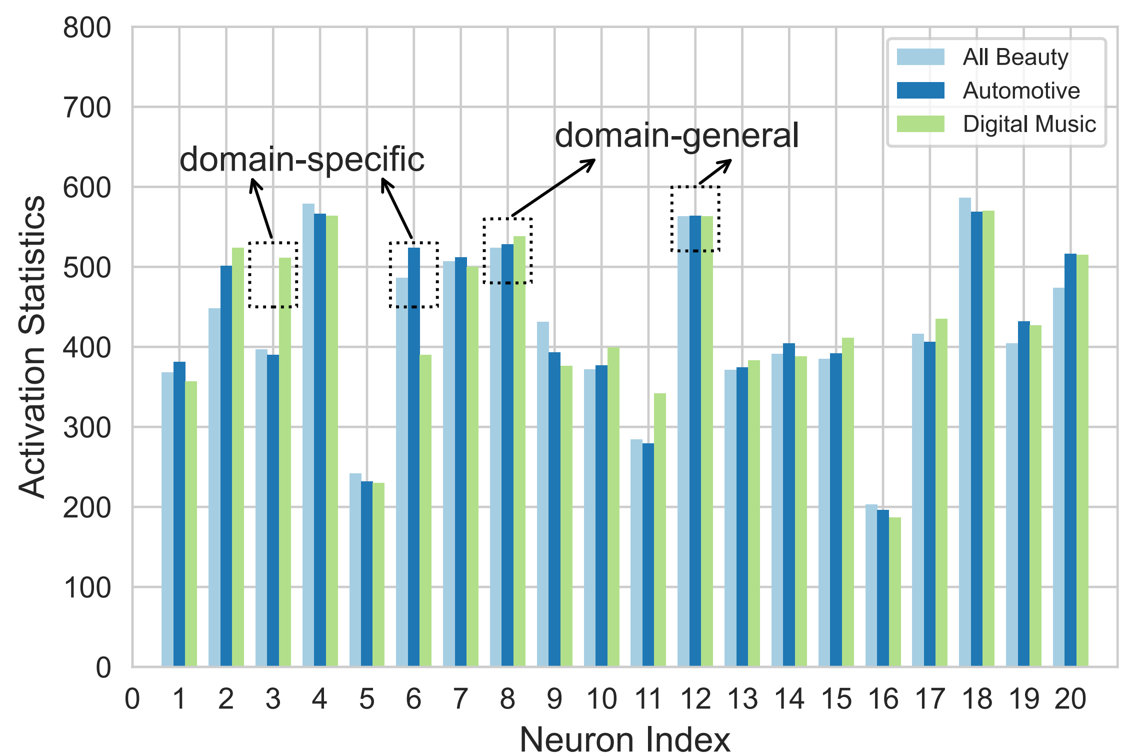
    - Given M training domains $D_{train} = \{D^i \,|\, i = 1,\ldots,M\}$ where $D^i = \{(x^i_j, y^i_j)\}_{j=1}^{n_i}$ denotes the i-th domain.

    - Domain generalization tends to learn a robust predictive function $h : X \rightarrow Y$ from D_{train} to achieve minimum prediction error on an unseen test domain D_{test}.

    - i.e., D_{test} cannot be accessed in training.

Motivation

- Over-parameterized LMs suffer from the limitation of large learning variance when faced with multiple domains.
- A pilot study on how different parameters of BERT behave over multiple domains.
- A critical portion of parameters are domain-specific while others are domain-general.



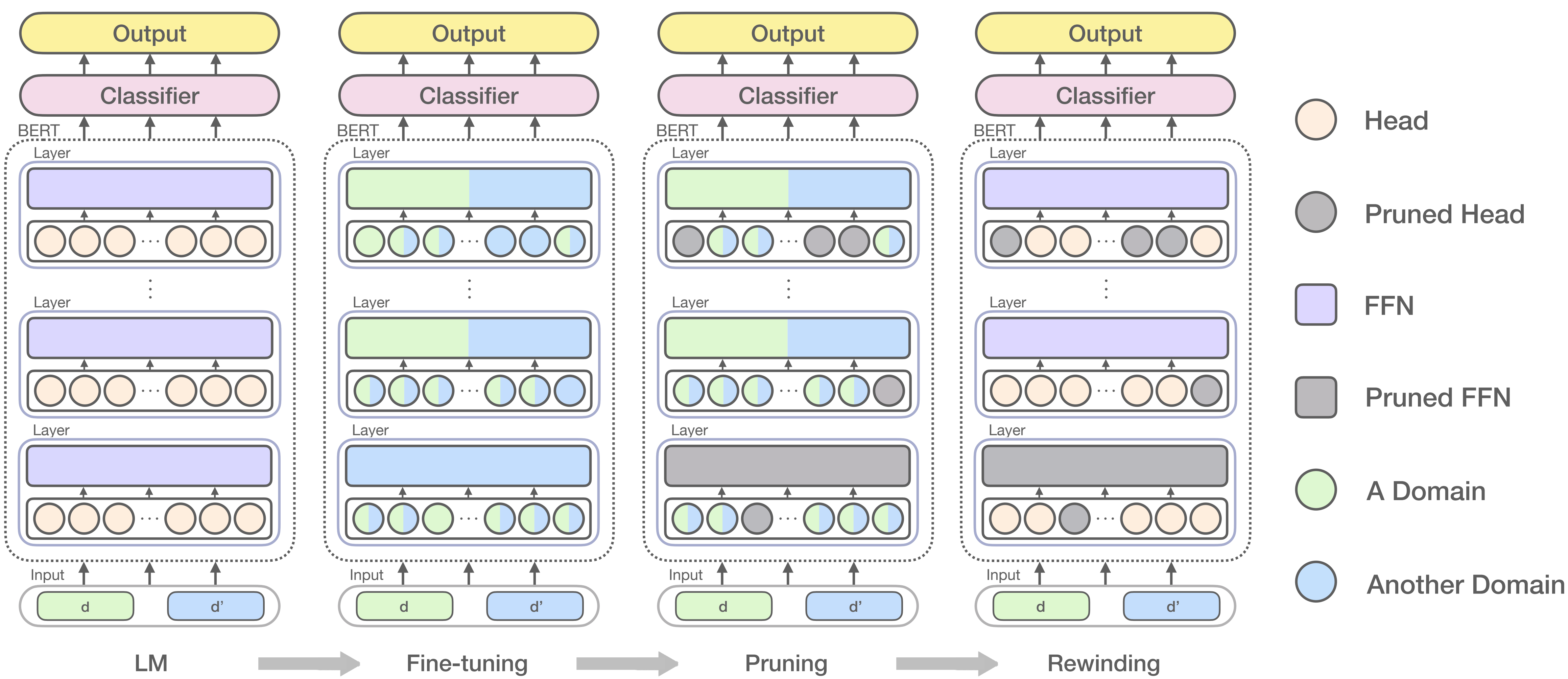
Motivation

- We posit that a domain-general LM underpinned by domain-general parameters can be derived from the original LM.
- The domain-general LM would facilitate a better domain generalization.
- Lottery tickets hypothesis states that a pruned model is capable of performing as expressive as the original over-parametrized model.
- We propose to identify domain-general parameters (dubbed *doge tickets*) by playing lottery tickets under the guidance of a domain-general score.

Method

- The identification of doge tickets follows a *first fine-tuning, then pruning, finally rewinding* paradigm.

- We apply structured pruning in LM by pruning MHA heads and FFN blocks.



Method

- Previous work identifies *winning tickets* by referring to the expressive scores of parameters.
- We approximate the expressive scores by masking elements of fine-tuned LM.

$$^{\circ}\mathrm{MHA}(\mathbf{X}) = \sum_{i=1}^{n} \xi^{(i)} \mathbf{H}^{(i)}(\mathbf{X}) \mathbf{W}_O^{(i)} \qquad\qquad ^{\circ}\mathrm{FFN}(\mathbf{Z}) = \nu \mathbf{W}_2 \mathrm{GELU}(\mathbf{W}_1 \mathbf{Z})$$

- **Expressive scores**

$$\mathbb{I}_{\mathrm{MHA}}^{(i)} = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left| \frac{\partial \mathcal{L}(x,y)}{\partial \xi^{(i)}} \right| \qquad\qquad \mathbb{I}_{\mathrm{FFN}} = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left| \frac{\partial \mathcal{L}(x,y)}{\partial \nu} \right|$$

- We propose a domain-general score which take the mean and variance of expressive scores across domains into account to identify the *doge tickets*.
- **Domain-general scores**

$$\begin{aligned}\mathbb{I}_{\mathrm{MHA}}^{(i)\prime} &= \mathbb{E}_{d\sim\mathcal{D}} \mathbb{E}_{(x,y)\sim d} \left| \frac{\partial \mathcal{L}(x,y)}{\partial \xi^{(i)}} \right| \\ &\quad - \lambda \mathbb{V}_{d\sim\mathcal{D}} \mathbb{E}_{(x,y)\sim d} \left| \frac{\partial \mathcal{L}(x,y)}{\partial \xi^{(i)}} \right|\end{aligned} \qquad\qquad \begin{aligned}\mathbb{I}_{\mathrm{FFN}}^{\prime} &= \mathbb{E}_{d\sim\mathcal{D}} \mathbb{E}_{(x,y)\sim d} \left| \frac{\partial \mathcal{L}(x,y)}{\partial \nu} \right| \\ &\quad - \lambda \mathbb{V}_{d\sim\mathcal{D}} \mathbb{E}_{(x,y)\sim d} \left| \frac{\partial \mathcal{L}(x,y)}{\partial \nu} \right|\end{aligned}$$

Experiments

- Out-of-domain datasets
  - The Amazon sentiment classification dataset
  - The MNLI language inference dataset
  - The OntoNotes named entity recognition dataset
- Baselines
  - BERT
  - BERT w. *IRM*
  - BERT w. *winning tickets*
  - BERT w. *random tickets*
  - BERT w. *doge tickets*

Dataset	$\mathcal{D}$	**#train.**	**#dev.**	$\mathcal{D}'$	**#test**
AMAZONA	{All Beauty, Automotive, Digital Music, Gift Cards}	5,400	600	{Industrial and Scientific, Movies, Software}	6,000
AMAZONB	{All Beauty, Industrial and Scientific, Movies, Software}			{Automotive, Digital Music, Gift Cards}	
AMAZONC	{Digital Music, Gift Cards, Movies, Software}			{All Beauty, Automotive, Industrial and Scientific}	
MNLI	{Fiction, Government, Slate, Telephone, Travel}	78,540	1,963	{Face to Face, Letters, Nine, Oup, Verbatim}	1,966
ONTONOTES	{Broadcast Conversation, Broadcast News, Magazine, Newswire}	16,111	2,488	{Telephone Conversation, Web Data}	1,837

Table 1: Statistics of datasets. **#train.**, **#dev.**, and **#test** indicate average number of training, development, and test examples per domain.

Results

- BERT w. *doge tickets* certainly generalizes better than baselines over all tasks.

Model	Datasets					Average Score	Average Sparsity
	AmazonA	AmazonB	AmazonC	Mnli	OntoNotes		
	Acc	**Acc**	**Acc**	**Acc**	**F1**		
BERT	69.8	72.6	69.6	84.8	57.2	70.8	0.0%
w/ IRM	70.4	72.5	70.7	84.3	56.3	70.8	0.0%
w/ *random tickets*	71.4	73.3	70.1	84.6	57.9	71.5	12.8%
w/ *winning tickets*	70.9	73.7	71.3	84.8	57.9	71.7	17.5%
w/ *doge tickets*	**71.7**	**73.8**	**72.2**	**85.0**	**58.5**	**72.2**	15.0%

Table 2: Main comparison results in percentage. The best results on datasets are **boldfaced**. Average Score is the average metric over used datasets. Average Sparsity is the average sparsity to achieve best out-of-domain generalization among all sparsity levels over used datasets.

Model	Datasets	Average Sparsity
	AmazonA	
	Acc	
BERT-large	73.1	0.0%
w/ IRM	73.5	0.0%
w/ *winning tickets*	74.0	15.0%
w/ *doge tickets*	**74.3**	15.0%

Table 3: Extended comparison results in percent. Larger LMs are used.

Analysis

- Sensitivity to Learning Variance

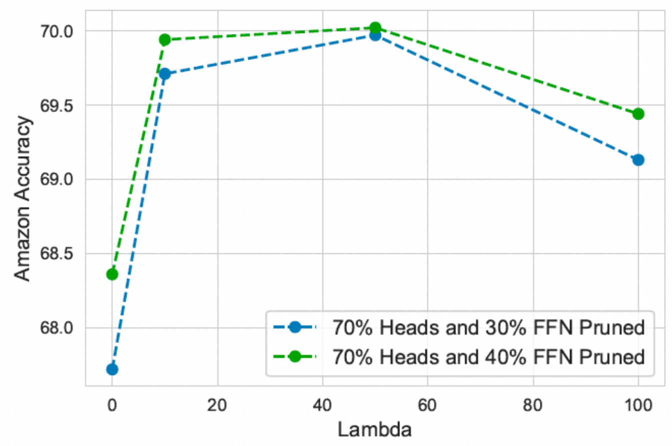


Figure 3: *doge tickets* on AMAZONA under various λ values with two sparsity levels.

- Impact of Training Domains
  - The impact of domain-specific (or domain-general) parameters on generalization becomes more significant.

Model	Datasets			Average Sparsity
	MNLI-5	MNLI-4	MNLI-3	
	Acc	**Acc**	**Acc**	
BERT	84.8	84.2	83.0	0.0%
w/ *winning tickets*	84.8	84.3	83.3	8.7%
w/ *doge tickets*	85.0	84.5	83.6	5.3%
Δ	0.2	0.3	0.6	–

Table 4: Results in percentage on MNLI with different training domain numbers. Δ means generalization margin.

Analysis

- Existence of Domain-specific Manner
  - High mean with high variance (HMHV)
  - High mean with low variance (HMLV)
  - Low mean with high variance (LMHV)
  - Ligh mean with low variance (LMLV)

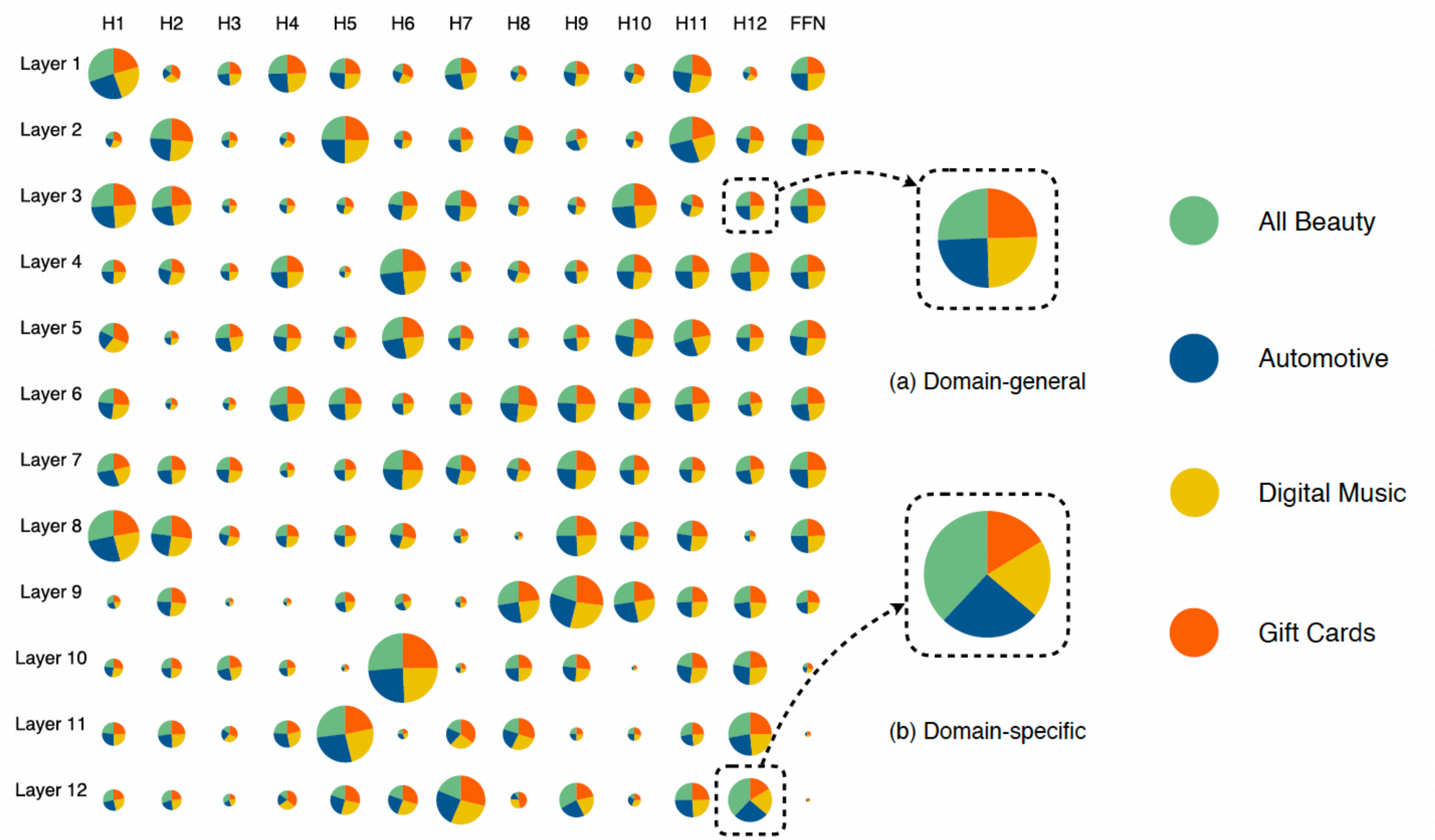
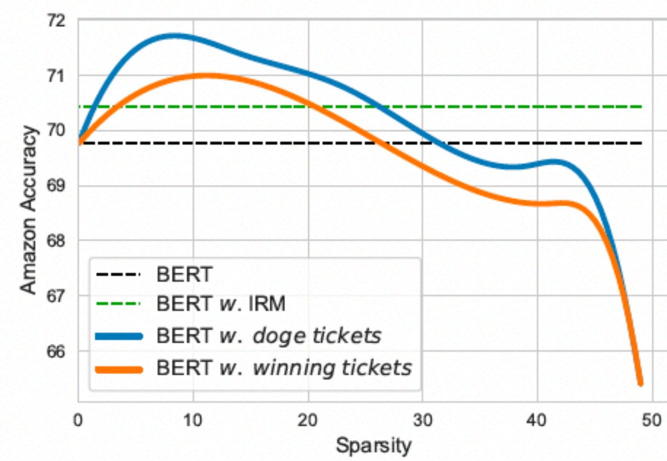


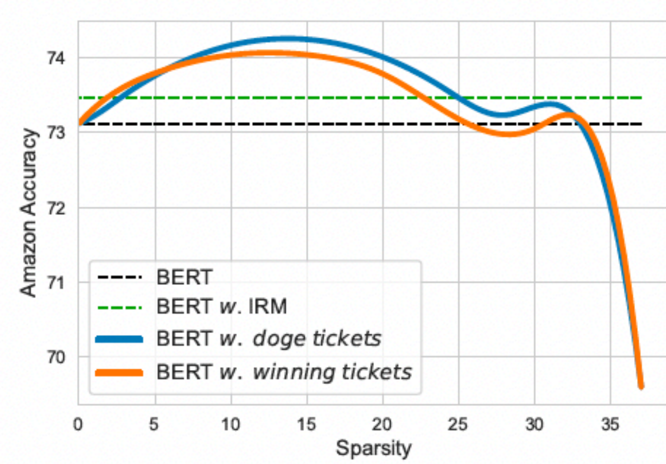
Figure 4: Illustration of expressive scores across domains. Each pie represents a parameterized element (either an MHA head or FFN block). The mean is measured by the radius of a pie. We use 4 distinguished colors to represent domains, whose details are shown in legend. The variance is measured by the proportion of each color in a pie.

Analysis

- Consistency with Varying Sparsity Levels
  - *Doge tickets* outperforms *winning tickets* most of the time.



(a) `BERT-base`



(b) `BERT-large`

Figure 5: Transitions with varying sparsity levels.

Conclusion

- We propose identify domain-general parameters by playing lottery tickets to uncover the domain-general LM.
- We propose a domain-general scores to guide the identification of *doge tickets*.
- *Doge tickets* shows advantages over previous *winning tickets* and the original over parameterized model on the out-of-domain datasets.
- arXiv: https://arxiv.org/abs/2207.09638
- Github: https://github.com/Ylily1015/DogeTickets

Thanks for your patience