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

Yi Yang[1], Chen Zhang[1], Benyou Wang[2], and Dawei Song[1(✉)]

[1] Beijing Institute of Technology, Beijing, China
{yang.yi,czhang,dwsong}@bit.edu.cn
[2] The Chinese University of Hong Kong, Shenzhen, China

Abstract. Over-parameterized pre-trained language models (LMs), have shown an appealing expressive power due to their small learning bias. However, the huge learning capacity of LMs can also lead to large learning variance. In a pilot study, we find that, when faced with multiple domains, a critical portion of parameters behave unexpectedly in a domain-specific manner while others behave in a domain-general one. Motivated by this phenomenon, we for the first time posit that domain-general parameters can underpin a domain-general LM that can be derived from the original LM. To uncover the domain-general LM, we propose to identify *do*main-*ge*neral parameters by playing lottery *tickets* (dubbed *doge tickets*). In order to intervene the lottery, we propose a *domain-general score*, which depicts how domain-invariant a parameter is by associating it with the variance. Comprehensive experiments are conducted on the AMAZON, MNLI, and ONTONOTES datasets. The results show that the *doge tickets* obtains an improved out-of-domain generalization in comparison with a range of competitive baselines. Analysis results further hint the existence of domain-general parameters and the performance consistency of *doge tickets*.

Keywords: Pre-trained language model · Domain generalization · Lottery tickets hypothesis

1 Introduction

It is witnessed that more and more models become increasingly over-parameterized, typically pre-trained language models (LMs). With the tremendous amounts of parameters, these models have shown an appealing expressive power [5,14,25,28]. While such LMs are enabled with small learning bias, they suffer from the limitation of large learning variance [20], especially when faced with multiple domains.

To show the consequence of such learning variance, we performed a pilot study on how different parameters of BERT behave over multiple domains (e.g., Digital Music, All Beauty, and Gift Cards). We find that a large portion

of parameters are domain-specific, while others are domain-general. Figure 1 shows an illustrative example. Certain neurons corresponding to domain-specific parameters may be particularly activated by some domains but not so active in others. The domain-inconsistency can potentially lead to a deteriorated out-of-domain generalization capability, concerning how well a LM performs universally on any domains [20,30]. In contrast, the neurons corresponding to domain-general parameters are similarly activated across different domains.

Motivated by the phenomenon, we for the first time posit that, a domain-general LM that is underpinned by domain-general parameters as illustrated in Fig. 1, can be derived from the original LM, and that the domain-general LM would facilitate a better out-of-domain generalization. Specifically, inspired by lottery ticket hypothesis stating that a pruned model is capable of performing as expressive as the original model, we propose to identify *do*main-*ge*neral parameters by playing lottery *tickets* (dubbed *doge tickets*) [6] under the guidance of a domain-general score. The domain-general score describes how domain-invariant a parameter is, rather than how expressive as depicted by the commonly used expressive score [18]. Driven by variance reduction [20], the domain-general score associates with the invariance of a parameter by looking at not only the mean but also the variance of its expressive scores across different domains.
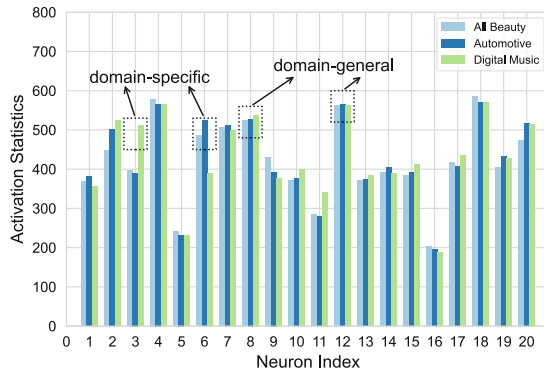
Fig. 1. A pilot study on how sampled neurons (for `[CLS]` token) corresponding to the last FFN block of the fine-tuned `BERT` [5] can be activated by training domains sampled from AMAZON [21] via activation statistics [2]. Considering that GELU [10] activation is used, we view neurons with values larger than minus two as activated. A parameter is said to be specific to a domain or a group of domains if its associated neurons are particularly activated by input from these domains, while not so active in other domains. A parameter is said to be general to domains if its associated neurons are similarly activated by input across different domains. Both domain-specific and domain-general manners are observed.

Comprehensive experiments are conducted on three datasets: the AMAZON [21], MNLI [29], and ONTONOTES [23] dataset. The results show that the proposed *doge tickets* owns a competitive out-of-domain generalization compared

with an array of state-of-the-art baselines. In-depth analyses are further carried out to double-check whether the domain-specific manner holds de facto and inspect whether the *doge tickets* performs consistently. The results hint the existence of domain-general parameters and the performance consistency of *doge tickets*.

2   Background

2.1   Out-of-domain Generalization

Pre-trained LMs [5,14,25] have achieved a compelling performance on downstream tasks where the training and test examples are identically distributed [28], thanks to the smaller learning bias brought by over-parameterization. Further exploration has also been carried out on whether LMs can have a good out-of-domain generalization [11], where the training and test examples are distinguished from each other in terms of distributions. However, the performance gain is somewhat limited or even degraded, due to the large learning variance brought by over-parameterization. Therefore, there is a large room for a better out-of-domain generalization of LMs [30].

   However, the issue with the large learning variance of pre-trained LMs is yet to be systematically explored. In a pilot study (c.f. Sect. 3.1), we find that under the setting of out-of-domain generalization, a number of parameters exhibit a domain-specific activation behavior, while others in a domain-general one. Motivated by this finding, we posit that domain-general parameters can underpin domain-general LMs. Therefore, provided that the training and test examples are separately sampled from training domains $\mathcal{D}$ and test domains $\mathcal{D}'$, we aim to uncover a domain-general LM $\mathcal{M}'$ from the original LM $\mathcal{M}$, and $\mathcal{M}'$ is expected to have a better out-of-domain generalization ability. It is also noteworthy that out-of-domain generalization is significantly different from domain adaptation, since domain adaptation has access to test examples.

2.2   Lottery Ticket Hypothesis

The lottery ticket hypothesis (LTH) states that a pruned model is capable of performing as good as the original over-parameterized model [6]. LTH can be applied to LMs in various ways [3,4,13,18] yet indicating a similar conclusion that *lottery tickets can win*. Previous work discovers *winning tickets* in LMs via either unstructured [4,6,26] or structured pruning techniques [4,18,24]. Unstructured pruning methods [8,16,22] concentrate on pruning parameters at neuron level. In contrast, structured pruning methods [9,12,17] prune parameters at module level. In particular, [13] investigates the transition behavior of *winning tickets* in LMs through shifting the sparsity levels, and has shown some slight performance gain with appropriate tickets (termed *super tickets*) at a small sparsity level. Inspired by LTH, we propose to identify the domain-general parameters by playing lottery tickets (dubbed *doge tickets*) under the guidance of a newly proposed domain-general score (c.f. Sect. 3.2).

2.3   Transformer Architecture

BERT [5] consists of a stack of transformer encoder layers [27]. Each of these layers contains two blocks: a multi-head self-attention block (MHA) and a feed-forward network block (FFN), with a residual connection and a normalization layer around each. In the following description, we omit the details of residual connections, normalization layers and bias terms for brevity. Assuming there are n independent heads in each MHA block and the i-th head is parameterized by $\mathbf{W}_Q^{(i)}$, $\mathbf{W}_K^{(i)}$, and $\mathbf{W}_V^{(i)} \in \mathbb{R}^{d \times d_h}$, then the output from the MHA block, which is sum of the intermediate outputs of all parallel heads, can be depicted as:

$$\mathbf{Z} = \mathrm{MHA}(\mathbf{X}) = \sum_{i=1}^n \mathbf{H}^{(i)}(\mathbf{X})\mathbf{W}_O^{(i)} = \sum_{i=1}^n \mathrm{Att}(\mathbf{X}\mathbf{W}_Q^{(i)}, \mathbf{X}\mathbf{W}_K^{(i)}, \mathbf{X}\mathbf{W}_V^{(i)})\mathbf{W}_O^{(i)}$$

where $\mathbf{X} \in \mathbb{R}^{l \times d}$ represents d-dimensional vectors that stand for l sequential input token representations and $\mathbf{W}_O^{(i)} \in \mathbb{R}^{d_h \times d}$ is the output linear layer for the i-th head. Likewise, we denote the output of the FFN block, which is composed of two linear layers with proper activation in-between as:

$$\mathbf{X}' = \mathrm{FFN}(\mathbf{Z}) = \mathbf{W}_2\mathrm{GELU}(\mathbf{W}_1\mathbf{Z})$$

where $\mathbf{W}_1 \in \mathbb{R}^{d \times d_i}$ and $\mathbf{W}_2 \in \mathbb{R}^{d_i \times d}$ correspond to the two layers. Our work exactly targets at improving out-of-domain generalization through uncovering a domain-general BERT from the original BERT by playing lottery tickets.

3   Identifying Doge Tickets

3.1   Uncovering Domain-general LM

LMs are perceived to bring large learning variance due to over-parameterization. Accordingly, we suspect that the variance to some extent showcases itself within
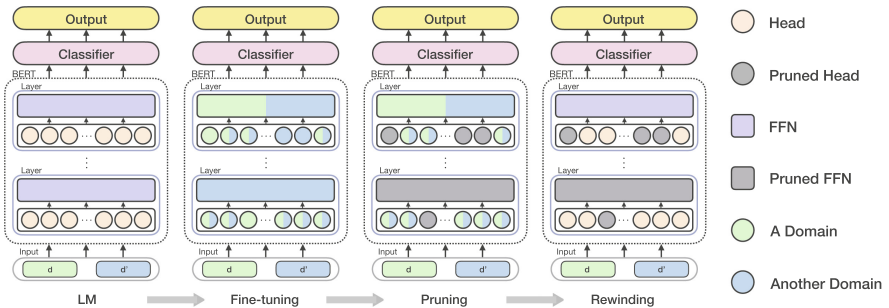
Fig. 2. *first fine-tuning, then pruning, finally rewinding* paradigm when playing lottery tickets. It is found that parameters can behave in either a domain-specific or domain-general manner after fine-tuning.

the LMs. Under the setting of out-of-domain generalization, we conduct a pilot study to examine how neurons activation [2] can vary from a domain to another, as in Fig. 1. In the study, the activation statistics towards three training domains of a sampled neuron (for [CLS] token) from each FFN block are attained by 1) feeding the examples sampled from these domains into the fined-tuned BERT; and 2) counting the accumulated times of activation. Recapping the results in Fig. 1, we can find that quite a few neurons, or parameters without loss of generality, unexpectedly behave in a domain-specific manner, which may well cause the unsatisfying out-of-domain generalization performance. Yet, other parameters may behave in a domain-general manner.

Motivated by the phenomenon, we hypothesize that domain-general parameters may essentially underpin a domain-general BERT that can be derived from the original BERT. This makes it possible to uncover a domain-general BERT by simply identifying domain-general parameters. Witnessing the success of LTH that states a well-performing pruned model can be uncovered by playing lottery tickets [6], we seek an effective solution customized for the domain-general BERT from intervening lottery tickets (dubbed *doge tickets*).

3.2   Playing Lottery Tickets

The identification of tickets typically follows a *first fine-tuning, then pruning, finally rewinding* paradigm [6]. Following recent advances [13,18,24], we would like to apply structured pruning to a LM to identify *doge tickets* by pruning MHA heads and FFN blocks. The most basic reason sits in that unstructured pruning requires maintaining a global pool of scores for all parameters, leading to a predominant memory overhead. Although we have only discovered domain-specific manner at neuron level in Sect. 3.1, we believe the same can apply at module level. The procedure is illustrated in Fig. 2, including the following steps: 1) fine-tuning: train the LM on a specific task; 2) pruning: prune parameters of the LM to a target sparsity under the guidance of the domain-general scores computed with the trained LM; 3) rewinding: set the pruned LM to the initial point of fine-tuning and train it on the task again.

Previous work identifies *winning tickets* by referring to the **expressive scores** of parameters [18]. While *winning tickets* can discover more light-weighted and more expressive LMs, we argue that they are not suitable to identify *doge tickets* since the variance has not been taken into consideration. Thus, we intervene the playing of lottery tickets guided by the **domain-general scores** which take into consideration the mean and variance of expressive scores across domains.

Expressive Scores. The expressive scores are approximated by masking the parameterized elements of the fine-tuned LM. $\xi^{(i)}$ and ν generally denote the mask variables respectively for an MHA head and an FFN block [24], such that:

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

where we initialize the mask variables $\xi^{(i)} = 1$ and $\nu = 1$ for the corresponding MHA head and FFN block to preserve the original LM. Then the expected absolute gradient over all training data for the MHA head and FFN block gives the expressive scores:

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

where (x, y) is a data point and $\mathcal{L}$ is the loss function. $\mathbb{E}$ represents expectation. The absolute value of gradient for a mask indicates how large the impact of masking the corresponding element is, thus implying how expressive an element is. Intuitively, a less expressive element should be pruned preferentially.

Domain-general Scores. After obtaining expressive scores in all domains, the domain-general scores take the mean and further the variance of expressive scores across domains. The domain-general scores can be formulated as:

$$\mathbb{I}_{\text{MHA}}^{(i)\prime} = \mathbb{E}_{d\sim\mathcal{D}}\mathbb{E}_{(x,y)\sim d} \left| \frac{\partial\mathcal{L}(x,y)}{\partial\xi^{(i)}} \right| \quad -\lambda\mathbb{V}_{d\sim\mathcal{D}}\mathbb{E}_{(x,y)\sim d} \left| \frac{\partial\mathcal{L}(x,y)}{\partial\xi^{(i)}} \right|,$$

$$\mathbb{I}_{\text{FFN}}' = \mathbb{E}_{d\sim\mathcal{D}}\mathbb{E}_{(x,y)\sim d} \left| \frac{\partial\mathcal{L}(x,y)}{\partial\nu} \right| \quad -\lambda\mathbb{V}_{d\sim\mathcal{D}}\mathbb{E}_{(x,y)\sim d} \left| \frac{\partial\mathcal{L}(x,y)}{\partial\nu} \right|$$

where (x, y) is a data point from the domain d. The domain-general score measures the balance between the mean and variance of expressive scores across domains. λ is adopted to quantify the trade-off.

As suggested by [19], we normalize the expressive scores of MHA heads with ℓ_2 norm. We gradually increase sparsity level by pruning the elements with lowest domain-general scores, therefore producing LMs with different sparsity levels. Rewinding is applied to these LMs by setting the remaining parameters' values to what they were in the early stage of fine-tuning [7] (i.e., the initialization point for fine-tuning). We expect the LM with remaining parameters can achieve better out-of-domain generalization, thus regarding them as *doge tickets*.

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

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

4   Experiments

4.1   Datasets

We conduct our experiments on the AMAZON sentiment classification dataset [21], the MNLI language inference dataset [29] and the ONTONOTES named entity recognition dataset [23]. They are described as follows:

- AMAZON originally contains reviews from a variety of product categories. We randomly select 7 product categories as domains and sample 6,000 reviews from each domain. Since each review is associated with a score ranging from 1 to 5 stars, we derive a 3-way sentiment labels from the score by viewing a score larger than 3 as positive, a score smaller than 3 as negative, and a score of 3 as neutral. Then we randomly select 4 out of 7 domains as training set ($\mathcal{D}$) and the rest 3 as test set ($\mathcal{D}'$). In this way, we can construct an out-of-domain dataset and repeat for three times, resulting in a total of 3 out-of-domain datasets, denoted as AMAZONA, AMAZONB, and AMAZONC respectively.
- MNLI covers a range of genres of spoken and written sentence pairs with textual entailment information. We strictly follow its original data split and use the mismatched (i.e., different genres from those in the training set) development set of the dataset as out-of-domain test set, giving us 5 training domains and 5 test domains.
- ONTONOTES consists of annotated named entities types from 6 genres. We randomly choose 4 domains as training set and the rest 2 as test set.

Table 1 present the detailed information and statistics of the datasets.

4.2   Models and Implementation

We directly fine-tune the pre-trained `BERT` [5] (in fact, `BERT-base`) without pruning to obtain a solid baseline for out-of-domain generalization measure, and further carry out comparisons among the following models:

- `BERT` w/ invariant risk minimization (`IRM`) applies a domain-invariant classifier [1] to the `BERT`, as a strong baseline for out-of-domain generalization.
- `BERT` w/ *random tickets* rewinds on a set of *random tickets* randomly sampled from the structured modules (MHA heads and FFN blocks).
- `BERT` w/ *winning tickets* rewinds on a set of *winning tickets* chosen according to the expressive scores over training data. We identify the tickets through pruning the `BERT` to an expected sparsity.
- `BERT` w/ *doge tickets* rewinds on a set of *doge tickets* chosen according to the domain-general scores over training data, which is akin to the procedure of `BERT` w/ *winning tickets*.

We use AdamW [15] as the optimizer in our experiments. The learning rate is searched within $\{1, 2, 3, 5\} \times 10^{-5}$ and the batch size is searched within $\{16, 32\}$. Both training and rewinding last no longer than 10 epochs, with

an early-stopping. For out-of-domain generalization, models are learned from training domains and evaluated on ever unseen test domains. The averaged Accuracy and the average F1-score over test domains are adopted as evaluation metrics during our experiments.

4.3   Main Comparison

Table 2 shows the main comparison results on all datasets. As we can see, `BERT` w/ *doge tickets* certainly generalizes better than `BERT`. Specifically, the improvements over `BERT` brought by *doge tickets* are 1.9% on AMAZONA, 1.2% on AMAZONB, 2.6% on AMAZONC, 0.2% on MNLI and 1.3% on ONTONOTES. Further, compared to the competitive out-of-domain generalization baseline, `BERT` w/ `IRM`, our model `BERT` w/ *doge tickets* also achieves a performance improvement, which illustrates that domain-specific parameters would better be pruned instead of regularized. Finally, `BERT` w/ *doge tickets* gains 0.5% on average over that with *winning tickets*, implying that domain variance should be considered as a significant part when playing the lotteries. Surprisingly, `BERT` w/ *random tickets* can be competitive with `BERT` w/ *winning tickets*, suggesting the sub-optimality of using *winning tickets* for out-of-domain generalization. Meanwhile, we notice that *doge tickets* favors a comparably smaller sparsity. This observation can be seen as an evidence that, with the variance, we should cautiously prune parameters. We also provide results for `BERT-large` in Table 3, which show that *doge tickets* can be applied to larger LMs and achieve a performance gain.

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

Model	Datasets
	AMAZONA
	Acc
BERT-large	73.1
w/ IRM	73.5
w/ *winning tickets*	74.0
w/ *doge tickets*	**74.3**

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

Model	Datasets		
	MNLI-5	MNLI-4	MNLI-3
	Acc	Acc	Acc
BERT	84.8	84.2	83.0
w/ *winning tickets*	84.8	84.3	83.3
w/ *doge tickets*	85.0	84.5	83.6
Δ	0.2	0.3	0.6

5   Analysis

5.1   Sensitivity to Learning Variance

To investigate the effect of learning variance for out-of-domain generalization, Fig. 3 shows the results from different settings, $\{0, 10, 50, 100\}$, of the variance coefficient λ on AMAZONA. The results suggest that it is necessary to take into account the variance when quantifying domain-general scores. The model performance first increases till a certain λ and then decreases, hinting that we can count the variance as much as possible on the same sparsity level to increase model generalization.
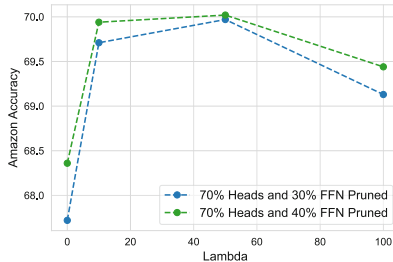

Fig. 3. *doge tickets* on AMAZONA under various λ values with two sparsity levels.

5.2   Impact of the Number of Training Domains

One may well note that the performance improvement of *doge tickets* is not significant on MNLI. We conjecture this is subject to the number of training domains. To explore whether training domain number matters, we conduct experiments by randomly selecting domains from the original 5 in MNLI (MNLI-5) as training domains, denoted as MNLI-4 and MNLI-3. As shown in Table 4, we observe that BERT w/ *doge tickets* discovers a larger generalization margin, when fewer (i.e., 4 and 3) training domains are used. That is, when using a smaller number of training domains, the impact of domain-specific (or domain-general) parameters on generalization results becomes more significant. According to the existing literature and intuition [11], more training domains can help LMs to extrapolate to unseen domains to some extent, thereby achieving a better out-of-domain generalization than the use of less training domains.

5.3   Existence of Domain-specific Manner

While we have found that a critical portion of parameters behave in a domain-specific manner through our pilot observation over intermediate activation distributions within the LM, we would like to verify the finding via intuitively visualizing expressive scores of parameters across domains. We use AMAZONA and show the expressive scores of elements across different domains in Fig. 4.
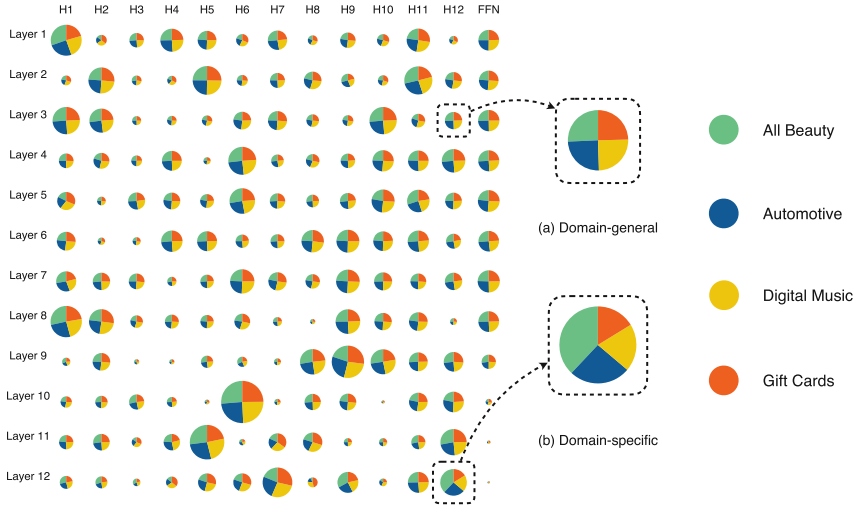
Fig. 4. Illustration of expressive scores across domains. Each pie represents a parameterized element (either an MHA head or FFN block). The mean and variance is measured by the radius of a pie and the proportion of each color in a pie respectively. We use 4 distinguished colors to represent domains, whose details are shown in legend.

Based on the mean and variance of expressive scores across domains, the parameters can be divided into 4 types: (1) high mean with high variance (HMHV); (2) high mean with low variance (HMLV); (3) low mean with high variance (LMHV); and (4) low mean with low variance (LMLV). It is obvious that there exists quite a number of elements with small radius, showing they play a less important role in the LM. This kind of parameters are exactly LMHV and LMLV parameters. If we merely consider this, *winning tickets* seems already a promising choice. However, more importantly, some elements are indeed behave unexpectedly in a domain-specific manner if we look at the proportions of different colors within pies, i.e., LMHV and HMHV parameters, suggesting the need of considering variance as in *doge tickets*.

To highlight, LMLV and HMLV elements are actually applicable across different domains, such as Fig. 4 (a). On the other hand, LMHV and HMHV ones are dedicated to a domain or two, for example "All Beauty" domain as Fig. 4 (b) shows. So *doge tickets* chooses to prune LMHV parameters in the first place yet HMLV in the last, leading to domain-general LMs.

5.4   Consistency with Varying Sparsity Levels

We prune the parameters under 25 sparsity levels for `BERT`. While both *doge tickets* and *winning tickets* can be identified at any sparsity levels, the evaluation results of them present transitions when we vary the sparsity levels continuously. We display the transitions of both `BERT-base` and `BERT-large` on Amazona in Fig. 5 to see whether *doge tickets* performs consistently better.
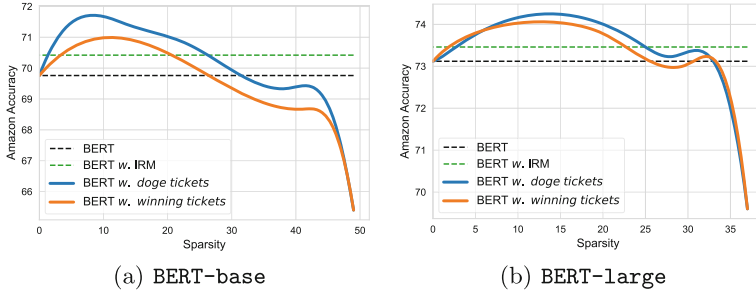
(a) `BERT-base` (b) `BERT-large`

Fig. 5. Transitions with varying sparsity levels.

We observe that *doge tickets* outperforms *winning tickets* usually. Recall the parameter types mentioned in Sect. 5.3. At smaller sparsity levels, *doge tickets* tends to first prune LMHV parameters (i.e., domain-specific parameters) instead of LMLV and LMHV parameters which are first considered by *winning tickets*. *doge tickets* shall preserve more domain-general parameters than winning tickets for a better out-of-domain generalization. On the other end, at larger sparsity levels, *winning tickets* and *doge tickets* start to prune expressive parameters. It occurs that HMLV parameters (i.e., domain-general parameters) will be pruned earlier than HMHV for *winning tickets*. Contrarily, *doge tickets* will certainly prune HMLV parameters later whenever possible. Hence, compared to *winning tickets*, *doge tickets* consistently makes LMs packed with invariant parameters at all sparsity levels to maintain a good out-of-domain generalization.

6 Conclusions

In this paper, we address the issue of learning variance of over-parameterized LMs, which we find through a pilot study can lead a critical portion of parameters to behave unexpectedly in a domain-specific manner. Neurons corresponding to these domain-specific parameters are particularly activated for some domains, yet not so active in others. Consequently, LMs are more likely to suffer from a low out-of-domain generalization, as it requires LMs to perform universally well on different domains.

Motivated by this observation, we posit that the other parameters are domain-general, essentially underpinning a domain-general LM that can appropriately derived from the original LM. To uncover the domain-general LM, we have proposed to identify domain-general parameters by playing lottery tickets under the vision of our proposed domain-general score (*doge tickets*). By taking into consideration both the mean and variance of parameter expressiveness, *doge tickets* shows advantages over previous *winning tickets* on the out-of-domain datasets. Further analyses verify the existence of domain-general parameters and performance consistency of *doge tickets*. We have empirically shown pruning can

improve out-of-domain generalization of LMs at large. In the future, we plan to examine the maximum potential of pruning by applying it under 1) unsupervised and 2) multi-task scenarios.

References

1. Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant risk minimization. arXiv:1907.02893 (2019)
2. Bengio, Y.: Deep learning of representations: looking forward. In: Dediu, A.-H., Martín-Vide, C., Mitkov, R., Truthe, B. (eds.) SLSP 2013. LNCS (LNAI), vol. 7978, pp. 1–37. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39593-2_1
3. Brix, C., Bahar, P., Ney, H.: Successfully applying the stabilized lottery ticket hypothesis to the transformer architecture. In: ACL, pp. 3909–3915 (2020)
4. Chen, T., et al.: The lottery ticket hypothesis for pre-trained BERT networks. In: NeurIPS (2020)
5. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL, pp. 4171–4186 (2019)
6. Frankle, J., Carbin, M.: The lottery ticket hypothesis: finding sparse, trainable neural networks. In: ICLR (2019)
7. Frankle, J., Dziugaite, G.K., Roy, D.M., Carbin, M.: Linear mode connectivity and the lottery ticket hypothesis. In: ICML. Proceedings of Machine Learning Research, vol. 119, pp. 3259–3269 (2020)
8. Han, S., Pool, J., Tran, J., Dally, W.J.: Learning both weights and connections for efficient neural network. In: NeurIPS, pp. 1135–1143 (2015)
9. He, Y., Zhang, X., Sun, J.: Channel pruning for accelerating very deep neural networks. In: ICCV, pp. 1398–1406 (2017)
10. Hendrycks, D., Gimpel, K.: Bridging nonlinearities and stochastic regularizers with gaussian error linear units. arXiv:1606.08415 (2016)
11. Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R., Song, D.: Pretrained transformers improve out-of-distribution robustness. In: ACL, pp. 2744–2751 (2020)
12. Li, H., Kadav, A., Durdanovic, I., Samet, H., Graf, H.P.: Pruning filters for efficient convnets. In: ICLR (2017)
13. Liang, C., et al.: Super tickets in pre-trained language models: from model compression to improving generalization. In: ACL, pp. 6524–6538 (2021)
14. Liu, Y., et al.: Roberta: a robustly optimized BERT pretraining approach. arXiv:1907.11692 (2019)
15. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
16. Louizos, C., Welling, M., Kingma, D.P.: Learning sparse neural networks through l_0 regularization. In: ICLR (2018)
17. Luo, J., Wu, J., Lin, W.: Thinet: A filter level pruning method for deep neural network compression. In: ICCV, pp. 5068–5076 (2017)
18. Michel, P., Levy, O., Neubig, G.: Are sixteen heads really better than one? In: NeurIPS, pp. 14014–14024 (2019)

19. Molchanov, P., Tyree, S., Karras, T., Aila, T., Kautz, J.: Pruning convolutional neural networks for resource efficient inference. In: ICLR (2017)
20. Namkoong, H., Duchi, J.C.: Variance-based regularization with convex objectives. In: NeurIPS, pp. 2971–2980 (2017)
21. Ni, J., Li, J., McAuley, J.J.: Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In: EMNLP, pp. 188–197 (2019)
22. Park, J., Li, S.R., Wen, W., Tang, P.T.P., Li, H., Chen, Y., Dubey, P.: Faster CNNs with direct sparse convolutions and guided pruning. In: ICLR (2017)
23. Pradhan, S., et al.: Towards robust linguistic analysis using ontonotes. In: Hockenmaier, J., Riedel, S. (eds.) CoNLL, pp. 143–152 (2013)
24. Prasanna, S., Rogers, A., Rumshisky, A.: When BERT plays the lottery, all tickets are winning. In: EMNLP, pp. 3208–3229 (2020)
25. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**, 140:1-140:67 (2020)
26. Renda, A., Frankle, J., Carbin, M.: Comparing rewinding and fine-tuning in neural network pruning. In: ICLR (2020)
27. Vaswani, A., et al.: Attention is all you need. In: NeurIPS, pp. 5998–6008 (2017)
28. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: GLUE: a multi-task benchmark and analysis platform for natural language understanding. In: ICLR (2019)
29. Williams, A., Nangia, N., Bowman, S.R.: A broad-coverage challenge corpus for sentence understanding through inference. In: NAACL, pp. 1112–1122 (2018)
30. Ye, H., Xie, C., Cai, T., Li, R., Li, Z., Wang, L.: Towards a theoretical framework of out-of-distribution generalization. arXiv:2106.04496 (2021)