ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет «Высшая школа экономики»

Факультет гуманитарных наук

Образовательная программа «Фундаментальная и компьютерная лингвистика»

КУРСОВАЯ РАБОТА

На тему «Хронологические подсети в модели BERT»

Тема на английском “Chronological subnetworks in BERT model”

Студентка 3 курса
группы № БКЛ-201
Ткач Анна Сергеевна

Научный руководитель
Сериков Олег Алексеевич
приглашенный
преподаватель Школы
лингвистики ФГН

Москва, 2023 г.

Contents

1. Introduction

Modern deep learning models have now achieved great progress in the field of language modeling and text generation. As a result, the pretrained language models (PLMs) are widely used for solving the majority of main NLP-problems (Howard and Ruder, 2018; Krause et al., 2019). It was shown by various researchers that state-of-the-art models, especially Transformer-based ones (such as BERT) are able to learn a significant number of language structures and language data during the training process.

Consequently, it would be interesting to research how these models acquire the language data from various points of view. In this work, we chronologically test the Lottery Ticket Hypothesis on a few checkpoints of the BERT model on various probing tasks.

Lottery Ticket Hypothesis (LTH) is the idea first introduced by J. Frankle and M. Carbin (Franklin and Carbin, 2018) which states that in any over-parameterized neural network there exists a subnetwork (a so-called *winning ticket*) which can achieve comparable results to the backbone architecture in a similar number of iterations.

In this work, we research the chronological subnetworks in BERT according to various language levels. For our study, we have used the probing method. We use probing tasks which reflect the three language levels: morphology, syntax and discourse. We do not look at phonology level, because we work with written texts.

We test six checkpoints of the pre-trained BERT model and for each task, we also check if there exists a subnetwork with performance comparable with the full model's one.

Our paper has the following structure: in the next chapter we observe the related literature and some essential theory. The third section of the paper discusses the methodology of our research: models we used, probing tasks and the methods of our study. Chapters 4 and 5 are dedicated to the results and the discussion, respectively. The last section is the conclusion. All the code and some other materials are available in the github repository[1].

[1] https://github.com/marianetta/Chronological_subnetworks

2. Related literature

2.1. Lottery Ticket Hypothesis & winning tickets

The LTH was first formulated in (Franklin and Carbin, 2018) and articulated that "dense, randomly-initialized, feed-forward networks contain subnetworks (*winning tickets)* that—when trained in isolation— reach test accuracy comparable to the original network in a similar number of iterations". Originally, the hypothesis was tested on fully-connected and convolutional networks.

Since then, there have been various studies that tested this hypothesis for other models. One of the first papers concerning winning tickets in Transformer models was (Yu et al., 2020). The authors studied LSTM models and Transformers and in both architectures they managed to find subnetworks which achieved nearly equivalent performance.

Then, there appeared plenty of studies dedicated to winning tickets in Pre-Trained Language Models (PLMs) and particularly BERT. One of the first papers on this field was (Chen 2020), which stated that in BERT, there can be found subnetworks (of size between 40% and 90% of the full model, in this study) which perform as well as a full model in case of retraining. In (Prasanna et al., 2020) it was shown that it is possible to find subnetworks that reach comparable performance with that of a full model, and that these subnetworks ("good" subnetworks) outperform others ("bad" subnetworks) of the same size.

In (Liang 2021) it was shown that, in the case of BERT, there are winning tickets that can even outperform the full model. (Gong et al., 2022) introduced the term *dominant winning ticket*, which is the subnetwork which can achieve satisfactory performance and can be found during the fine-tuning of the PLM. These dominant tickets have the following features: they can achieve the performance comparable with the full model's one; they are transferable across various tasks; they are highly structured (these subnetworks resemble the "skeleton" of the backbone network). The researchers claim that in case of RoBERTa-large there exist efficient subnetworks which have the size less than 0.2% compared to the whole model.

2.2. Language levels and language models

It is a common theory in linguistics that language operates as a system of five levels: phonology, morphology, syntax, semantics and discourse. Moreover, these levels are considered to function consistently (for example, phonological units form morphemes, etc.).

There have been various studies which tried to provide insight into the process of language acquisition during models' training. For instance, in (Belinkov et al., 2017) it is stated that lower layers of Neural Machine Translation encoders have more information about morphological features, while higher layers contain more syntactic information. (Rogers et al., 2020) surveyed the localization of linguistic knowledge in BERT: the most information about word order is contained in lower levels, syntactic information is localized in middle levels and semantics is spread across different levels of the model.

2.3. Probing approach

Since we have used a probing method in our study, it is also needed to overview this approach.

Probing tasks were first introduced in (Conneau et al., 2018). In this work, the most popular probing approach, diagnostic classifiers, is described. The authors articulate that a probing task is a "classification problem that focuses on simple linguistic properties of sentences". The idea of probing tasks is the following: researchers take the sentence representations from the model they study and input them into a probing classifier; then, the task for the classifier is to assign a label of some linguistic feature to each of its input representations. For the sake of minimizing interpretability problems, probing tasks should ask a simple question (for example, a subject's number).

However, there are also other probing methods. One of the most comprehensive classification of probing approaches can be found in (Belinkov et al., 2020). It is said that there are two main categories of probing methods: structural analysis (the method we use in our study belongs to this category) and behavioral studies. The behavioral methods do not require any classifier.

The probing approach can be used chronologically. This approach is described, for instance, in (Voloshina et al., 2022). The idea is to run experiments with probing tasks on different checkpoints (different number of training iterations) of the model and compare the model's performance on them.

In our study we combine the ideas of winning tickets and chronological probing on different language levels tasks. On six checkpoints of pre-trained BERT, we attempt to find efficient subnetworks, whose performance on different tasks is comparable with that of the full model.

3. Methodology

3.1. Models

At first, we conducted a series of experiments with a base version of BERT (bert-base-uncased from the transformers library by HuggingFace) as a baseline. We have used the open probing framework introduced by AIRI-institute[2].

Then, to perform our main series of experiments, we needed the trained BERT. We have trained the BERT model, using the BERT-mini config[3]. We have trained it on book-corpus and wikipedia datasets by HuggingFace, saving 6 checkpoints from 100.000 steps to 1.000.000 steps (100000, 200000, 400000, 600000, 800000, 1000000). These checkpoints we have used later in our experiments.

3.2. Probing Tasks

For our research, we used probing tasks from two datasets: SentEval[4] (morphology and syntax tasks) and DiscoEval[5] (discourse tasks). SentEval was first described in (Conneau et al., 2018). The paper (Chen et al., 2019) is dedicated to the DiscoEval dataset. We have taken two tasks for each language level. As a result, in total, during the experiments we have used six different tasks, which are described below:

[2] https://github.com/AIRI-Institute/Probing_framework
[3] https://github.com/google-research/bert
[4] https://github.com/facebookresearch/SentEval
[5] https://github.com/ZeweiChu/DiscoEval

1. Subject number (SN): this task shows how well the model has learned to identify the sentence subject's number. This is a binary classification task with labels NN and NNS which stand for singular and plural number, respectively.
2. Object number (ON): this is a task to determine the number of the sentence's object. ON, as well as SN, is a binary classification task with labels NN and NNS which stand for singular and plural number.
3. Bigram Shift (BS): the task, which allows to define, if the model is sensitive to legal word orders. This dataset contains original sentences and their pairs – sentences, where two random adjacent words are inverted. This task represents a binary classification task with labels O (original) and I (inverted).
4. Tree Depth (TD): the task shows if the model is capable of learning to define the depth of the syntactic tree of the sentence. This is a multiclass classification task, the depth of trees in the dataset varies from 5 to 12.
5. Discourse coherence (DC): this is the task to define, if six sequential sentences constitute a coherent paragraph. This is a binary classification task with labels 0 (incoherent text) and 1 (coherent text).
6. Sentence position (SP): the dataset contains sequences from five sentences, one of which is transposed to the first place. The goal of the model is to determine the initial position of the transposed sentence.

In the above list, the first two tasks reflect the morphology level, tasks 3 and 4 stand for syntax level, and the last two ones are for discourse level. The examples of tasks are presented in Table 1.

3.3. Method

As a baseline, we have conducted a series of experiments with a base BERT model (namely, bert-base-uncased from the transformers library by HuggingFace). We have run six baseline experiments, one for each of the tasks.

At the main stage of our study, we have conducted experiments with the BERT model, training which is described in section 3.1.

To find and access the subnetworks in our BERT, we have used the NeuroX library[6]. We have accessed the performance of subnetworks of size 10% and 25% (by

[6] https://github.com/fdalvi/NeuroX

leaving top-n neurons of the model) of the full model's size for each BERT's checkpoint on each task.

Task	Examples	Labels
Subject number	Chills ran along her skin just thinking about the terrifying experience. The van pulled up to the church and everyone got out.	NNS NN
Object number	He signaled the bartender and ordered one more drink. The Afghans had suffered heavy losses.	NN NNS
Bigram shift	I hated even hearing that name now. This is my Eve Christmas.	O I
Tree depth	Who knew who would be there ? He jumped as a coarse whisper broke his concentration.	10 8
Discourse coherence	<user> : same thing with my girlfriend . damn i think my sea <unka> died i can only relate what i know it just does n't work :-) <user> : same here . that may be true in your case <user> : same thing with my girlfriend . i do n't know your particular circumstances i can only relate what i know it just does n't work :-) <user> : same here . that may be true in your case	0 1
Sentence position	Dan was overweight as well. Dan 's parents were overweight. The doctors told his parents it was unhealthy . His parents understood and decided to make a change. They got themselves and Dan on a diet. Jane did n't know how to react . Jane was working at a diner . Suddenly , a customer barged up to the counter . He began yelling about how long his food was taking . Luckily , her coworker intervened and calmed the man down.	1 3

Table 1. Task examples

4. Results

4.1. Baseline experiments (Bert-based)

As a baseline, we have measured the performance of bert-base-uncased from Transformers library on the six tasks. The score we have used is accuracy. The results are shown in Figure 1 below.

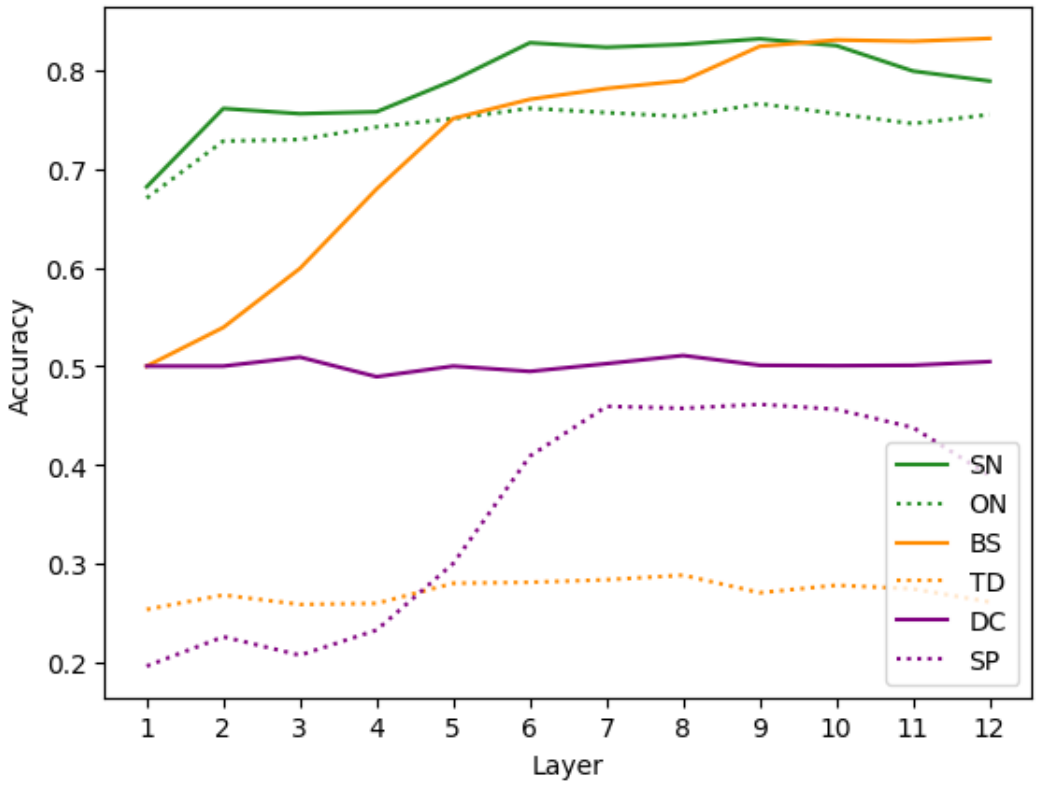

Figure 1. Accuracy of bert-base-uncased on the six tasks.

As can be seen, morphology tasks are the easiest for the model. The model's accuracy on Subject number and Object number varies from 0.67 to 0.83.

The performance of the model in case of syntax is different for two tasks. The accuracy on Bigram shift increases from 0.5 to more than 0.8, while the accuracy on Tree depth task fluctuates around 0.3.

The performance on discourse tasks is also rather poor: approximately 0.5 for Discourse coherence and between 0.1 and 0.44 for Sentence position.

4.2. Experiments with trained BERT-mini

The resulting accuracy score of our BERT and its subnetworks on various probing tasks is presented below, on Figures 2-7 (*full* stands for the full BERT, *sub10* – for the subnetwork of size 10% of the full model's size, *sub25* – for the subnetwork of size 25% of the full model's size).

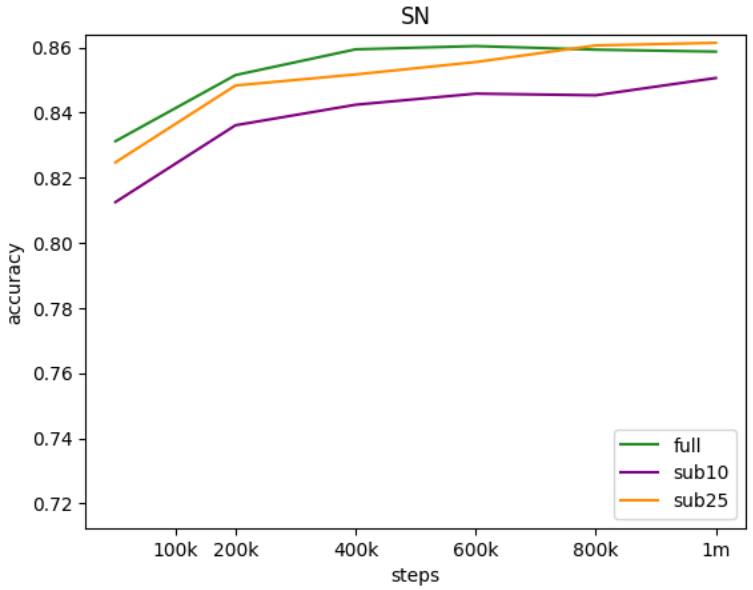

Figure 2. Accuracy of trained BERT and subnetworks on the SN task.

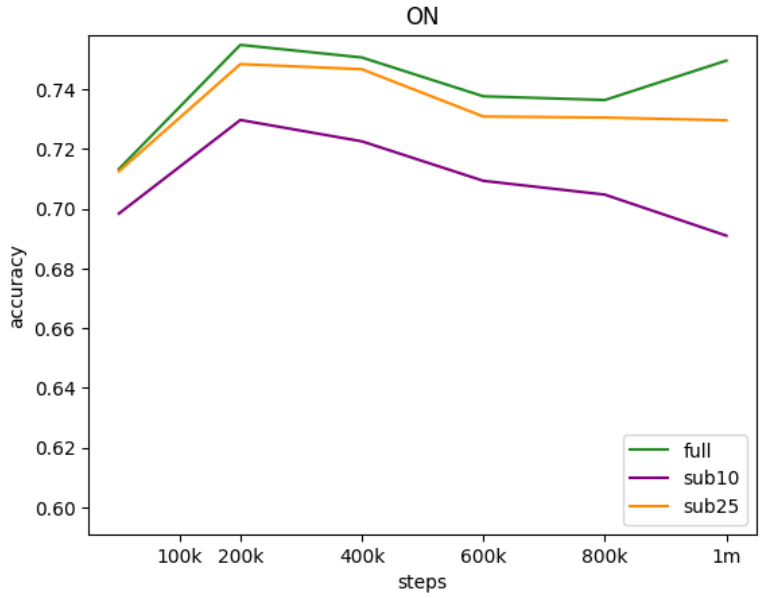

Figure 3. Accuracy of trained BERT and subnetworks on the ON task.

As can be seen, the accuracy of the full BERT on morphology tasks is close to that of the baseline model: for Subject number, it is from 0.8 to 0.86 and for Object number it is from 0.7 to 0.75. The accuracy curves for both subnetworks look similar to the backbone model and their score is quite high, comparable with that of the full BERT. For the SN task, the best performance of the full BERT is on the checkpoint-400000 and for both subnetworks it is on the last checkpoint. For the ON task, the best performance of all three models is on the checkpoint-200000.

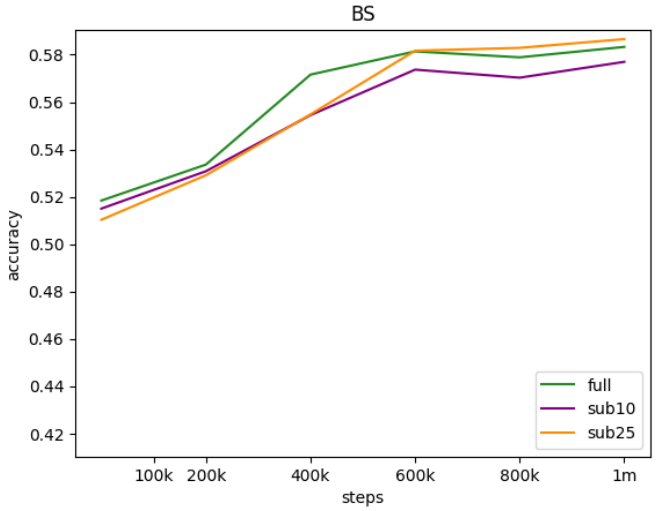

Figure 4. Accuracy of trained BERT and subnetworks on the BS task.

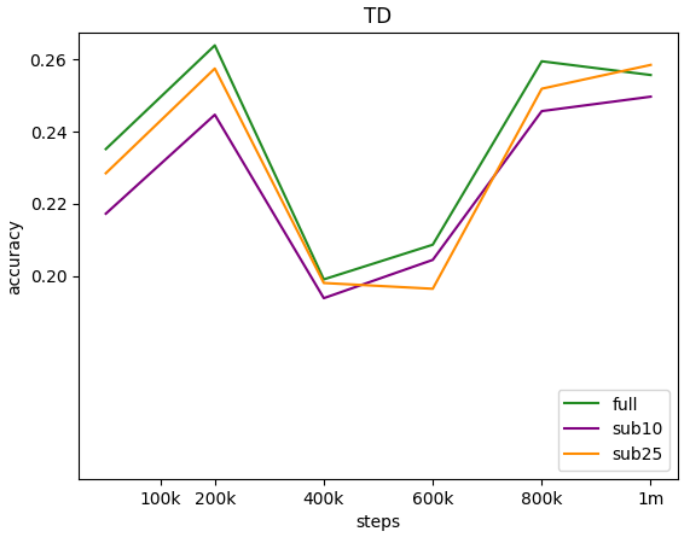

Figure 5. Accuracy of trained BERT and subnetworks on the TD task.

The accuracy on bigram shift task contradicts with the baseline results – it is rather low for all three models. All three models achieve the highest score on the last checkpoint. Concerning the tree depth task, the accuracy curves look similar for three models: the score is the highest on the checkpoint-200000 and the last two checkpoints.

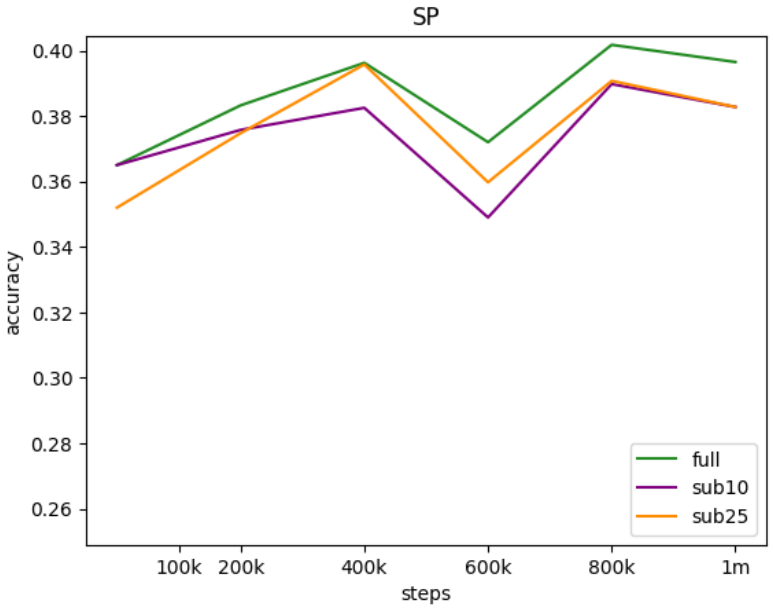

Figure 6. Accuracy of trained BERT and subnetworks on the SP task.

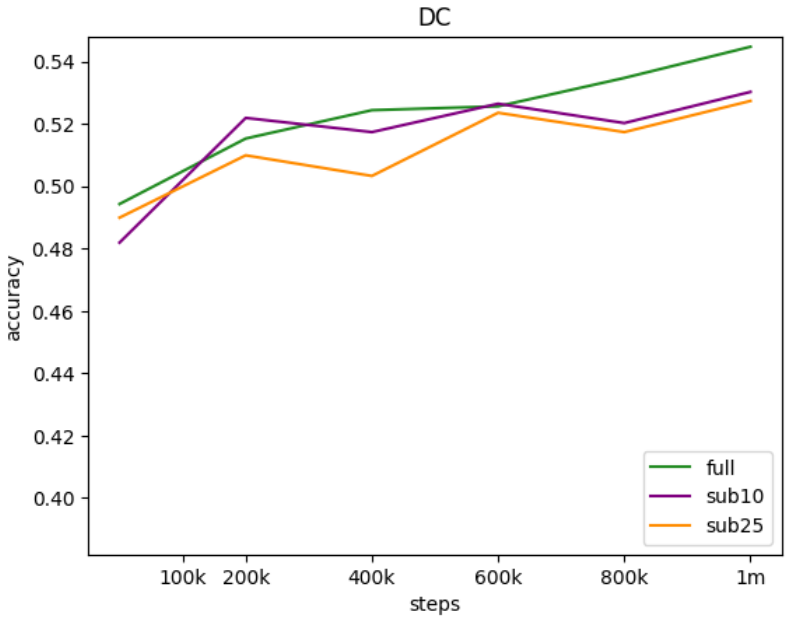

Figure 7. Accuracy of trained BERT and subnetworks on the DC task.

The score of models on discourse tasks look similar to the baseline ones, the accuracy is only slightly higher for discourse coherence tasks for the last checkpoints of BERT. The subnetworks here again perform almost the same as the full model. It is common for three models to reach the highest accuracy on the last checkpoint, but the smaller subnetwork performs the best on DC on checkpoint-200000.

5. Discussion

5.1 Morphology tasks

As our results show, the subnetworks of our BERT reach a performance which is slightly worse than the performance of the full model from the first checkpoint to the last. Furthermore, the performance of the bigger subnetworks (of size 25%) is moderately better than that of the smaller one. Although it is typical for BERT to achieve the best performance on morphology tasks on the middle checkpoints, our subnetworks do not always follow this pattern (their best accuracy on the SN task is on the last checkpoint).

5.2 Syntax tasks

Concerning syntax probing tasks, our results are rather unclear. The accuracy of our trained BERT (and its subnetworks) on BS is lower than the baseline score for some reasons. However, the subnetworks reach the comparable score again, and the bigger one achieves slightly better results than the smaller one. The performance on syntax tasks is high on the last checkpoints for all models, but for TD it is also high on checkpoint-200000. The bigger subnetwork is generally better for syntax tasks.

5.3 Discourse tasks

In terms of discourse tasks, both subnetworks perform almost the same as the full model, too. For the sentence position task, the bigger subnetwork is slightly more effective than the smaller one, starting from checkpoint-200000. For the discourse coherence task, the smaller subnetwork is better (and it even outperforms the full model

on checkpoint-200000 and checkpoint-600000). However, the score of the model on DC is quite low. The best performance of all three models is most often on the last checkpoints (800000 or 1000000).

Overall, it is seen that the existence of efficient subnetworks does not vary for different language levels. It is also clear that the performance of these subnetworks is close to the full model's one on all checkpoints, starting from the very first one.

6. Conclusion

This paper addresses the issue of efficient subnetworks (winning tickets) in BERT from chronological point of view and also in the perspective of different language levels tasks. In our study we tried to find subnetworks of size 10% and 25% of the full model's size which have comparable performance with that of the full model.

We have conducted experiments with six checkpoints of trained BERT on six probing tasks. Our results show that subnetworks of top-n neurons demonstrate the performance which is near that of the full BERT on every checkpoint. This pattern is also the same for morphology, syntax and discourse tasks.

However, the performance of our trained BERT, especially on discourse tasks, is not good enough. In the case of discourse, it could be because of the smaller size of task datasets. It could be useful to estimate the model and its subnetworks on a wider dataset in future research. Also, it would be profitable to try assessing more checkpoints, probing different learning parameters or using other metrics (as f1-score).

References

Belinkov, Y., Marquez, L., Sajjad, H., Durrani, N., Dalvi, F., Glass, J. (2017).Evaluating Layers of Representation in Neural Machine Translation on Part-of-Speech and Semantic Tagging Tasks. *Proceedings of the The 8th International Joint Conference on Natural Language Processing*, 1–10.

Belinkov, Y., Gehrmann, S., & Pavlick, E. (2020). Interpretability and analysis in neural nlp. *Proceedings of the 58th annual meeting of the association for computational linguistics: tutorial abstracts*, 1–5.

Chen, M., Chu, Z., & Gimpel, K. (2019). Evaluation benchmarks and learning criteria for discourse-aware sentence representations. *Proc. of EMNLP.*

Chen, T., Frankle, J., Chang, S., Liu, S.,Zhang, Y., Wang, Z., & Carbin, M. (2020). The lottery ticket hypothesis for pre-trained bert networks. *arXiv preprint arXiv:2007.12223.*

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., & Baroni, M. (2018). What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070.*

Frankle, J. & Carbin, M. (2018). The lottery ticket hypothesis: Finding sparse, trainable neural networks. *In International Conference on Learning Representations, 2018.*

Gong, Z., He, D., Shen, Y., Liu, T., Chen, W., Zhao, D., Wen, J., & Yan, R. (2022). Finding the Dominant Winning Ticket in Pre-Trained Language Models. *Findings of the Association for Computational Linguistics: ACL 2022,* 1459-1472.

Howard, J., Ruder, S. (2018). Universal language model fine-tuning for text classification. *In ACL. Association for Computational Linguistics.*

Krause, B., Kahembwe, E., Murray, I., & Renals, S. (2019). Dynamic evaluation of transformer language models. *arXiv preprint arXiv:1904.08378.*

Liang, C., Zuo, S., Chen, M., Jiang, H., Liu, H., He, P., Zhao, T., & Chen, W. (2021). Super tickets in pre-trained language models: From model compression to improving generalization. *arXiv preprint arXiv:2105.12002.*

Prasanna, S., Rogers, A., & Rumshisky, A. (2020). When bert plays the lottery, all tickets are winning. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP),* 3208–3229.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics,* 8, 842–866.

Voloshina, E., Serikov, O., Shavrina, T. (2022). Is neural language acquisition similar to natural? A chronological probing study. *arXiv preprint arXiv:2207.00560.*

Yu, H., Edunov, S., Tian, & Y., Morcos, A. (2020). Playing the lottery with rewards and multiple languages: Lottery tickets in RL and NLP. *In ICLR.*