Critical Size Hypothesis: How Model Hyperparameters Correlate with Its Linguistic Abilities

Ekaterina Voloshina
University of Gothenburg
Chalmers University of Technology
ekaterina.voloshina@chalmers.se

Oleg Serikov
KAUST
oleg.serikov@kaust.edu.sa

Abstract

In recent years, the models were tested on different probing tasks to examine their language knowledge. However, few researchers explored the very process of models' language acquisition. Nevertheless, the analysis of language acquisition during training could shed light on the model parameters that help to acquire the language faster. In this work, we show how the model architecture seems not to influence the language acquisition process. We experiment with model hyperparameters and reveal that the hidden size is the most essential factor for model language acquisition.

1 Introduction

Modern deep learning models have achieved significant results in the field of language modeling and text generation (Krause et al., 2019; Niu et al., 2020). Therefore, language models (LMs) are often used in linguistic research to find systematic similarities in the language data. Performance of the state-of-the-art models, such as Transformer-based ones (Vaswani et al., 2017), on linguistic tasks show that they have learned measurable language structures during the training process (Warstadt and Bowman, 2022).

Consequently, it is interesting to explore how the LMs acquire the language during their training process and what part of their architecture helps to acquire a language better. In this work, we study the correlation between the acquisition process in the BERT model and different model sizes. Linguistic tasks are meant to represent three levels of language grammar structure: morphology, syntax, and discourse. In other words, we pose the following questions: which parameters of models influence the language acquisition process?

2 Related work

The first work on probing of neural networks across time was carried by Saphra and Lopez (2018). The authors showed that first, a LSTM model (Hochreiter and Schmidhuber, 1997) acquires syntactic and semantic features and later information structure. Chiang et al. (2020) looked at the training process of ALBERT (Lan et al., 2019) and concluded that semantic and syntactic information is acquired during the early steps while accuracy on world knowledge fluctuates during the training. Liu et al. (2021) showed similar results on RoBERTa (Liu et al., 2019): the model shows good results on linguistic probing tasks starting from early stages, and later it learns factual and common sense knowledge. (Blevins et al., 2022) studied training dynamics of multilingual models, they reveal that while linguistic information is acquired early, transfer learning abilities are evolving during the entire training process. Choshen et al. (2022) examined the trajectories of models' language acquisition, and they find no impact of architecture or a model size on training trajectories. Warstadt and Bowman (2022) provides survey and theoretical discussions on how neural networks can help us learn more about language acquisition. Following one of the ideas we conduct an ablation study of model's hyperparameters.

3 Methods

3.1 Models

We train small models to see how language acquisition trajectories vary depending on the model hyperparameters. Since previous research shows that the acquisition of most of the linguistic features stops after 500,000 steps, we look at the first training steps. We regard number of layers, embedding size and number of attention heads to be crucial. Therefore we train four models:

1. A base model: the hidden size of 128, 2 layers, and 2 attention heads;
2. A model with increased number of attention heads: the hidden size of 128, 2 layers, and 4 attention heads;

Proceedings of the 2024 CLASP Conference on Multimodality and Interaction in Language Learning, pages 1–7

3. A model with increased hidden size: the hidden size of 256, 2 layers, and 2 attention heads;
4. A model with increased number of layers: the hidden size of 128, 4 layers, and 2 attention heads.

Our hypothesis states that if any of these models show a significantly different result on any group of tasks, this parameter causes a better acquisition process. If all the models show similar results, different sizes of models do not correlate with the acquisition process, therefore, it depends on language features rather than on model parameters.

We train models with the same computational resources and data corpus, which included Wikipedia articles limited to 10,000,000 tokens. We choose this threshold as an optimal one, as according to Zhang et al. (2020), models can acquire basic linguistic information from this amount of data. We compare our model to **MultiBERT** (Sellam et al., 2021), the model with 12 layers and embedding size 768. Unlike the original BERT (Devlin et al., 2018), it was trained with 25 different seeds. We use the model with seed 0 and we use the same seed to train small models to make our results more comparable.

To explore the combination of different hyperparameters, we train several other models. We are interested in what size the model should have to behave as the model of standard size (768 embedding size, 12 layers and 12 attention heads). To calculate that, we first train a model of the same size as the multiBERT we compared to in the experiments before. Then we use it as a standard of comparison and train several models of different sizes on the same data and with the same setup as the standard BERT. We limit the training process to 100,000 iterations to find minimal parameters that help the model to achieve the accuracy of the standard model. Table 1 summarise models we trained to find the proper combination of parameters.

Model	**Size**	**Layers**	**Att. heads**
1	256	4	4
2	256	8	4
3	512	4	4
4	512	8	8
5	512	12	8
6	768	8	8
7	768	12	8

Table 1: Summarisation of trained models: for each model we state the hidden size of embeddings, number of layers, and number of attention heads.

3.2 Probing tasks

We use probing tasks from several probing datasets, such as SentEval (Conneau et al., 2018), Morph Call (Mikhailov et al., 2021), DisSent (Nie et al., 2019), DiscoEval (Chen et al., 2019), and BLiMP (Warstadt et al., 2020) (see examples in Tables 3 and 4):

- **Transitive verbs** includes minimal pairs of sentences with different verbs, where only one verb is transitive.
- **Passive verbs** consists of pairs that have different verbs, where only one verb can be used in a passive form.
- **Island effects** tests a model's sensibility to syntactic order. An island is a structure from which a word cannot be moved (Ross, 1967).
- **Principle A** shows the use of reflexives. According to Chomsky (1981), a reflexive should have a local antecedent, and if it does not, the sentence is ungrammatical.
- **Subject number** is a binary classification task with labels NNS and NN (plural and singular number, respectively).
- **Person** is a binary classification with labels 0 and 1, which signifies if a subject has a person marker or not.
- **Tree depth** contains six classes, each of which stand for a depth of the syntactic tree of a given sentence.
- **Top constituents** requires to identify the number of constituents located right below the sentence (S) node.
- **Connectors** includes pairs of sentences originally connected with one of 5 prepositions, and the task is to choose the omitted preposition.
- **Sentence position** contains sequences of 5 sentences, and the first sentence is placed in the wrong place. Therefore, the aim is to detect the original position of these sentences.
- **Penn Discourse Treebank** is based on Penn Discourse Treebank annotation (Marcus et al., 1994). The aim is to choose the right discourse relation between two discourse items from Penn Treebank.
- **Discourse coherence** is a binary classification with classes 1 and 0. Class 1 means that

the given paragraph is coherent, and class 0 should be assigned to paragraphs with shuffled sentences.

For tasks from BLiMP, we mask each word in a sentence, and sum probabilities of all words. The probability of an acceptable sentence should be higher than the probability of an unacceptable sentence.

For most tasks we take a sentence embedding via mean pooling. A logistic regression as a classifier model is used to classify embedded sentences. For the Sentence Position task, we calculate the difference between the first embedding and the other pairwise. The first embedding and its differences with others are concatenated and put as an input to a classifier. For other discourse tasks, we concatenated sentence embeddings, which were calculated as the mean of token embeddings.

4 Results and Discussion

First, we conducted the same experiments on four small models described in section 3.1.

As Figure 1 shows, compared to MultiBERT, models show worse accuracy. However, among small models, the one with the increased hidden size shows the best results in all cases, except for Penn Discourse Treebank and Tree depth, where the model with the increased number of layers shows the best results. This model shows the second best results on other tasks.

The behaviour of the model with the increased number of attention heads is inconsistent compared to the *base* model (hidden size of 128, 2 attention heads, and 2 layers). On some tasks, such as Penn Discourse Treebank and Discourse Coherence, it shows worse accuracy than the base model. On other tasks, it shows better quality than the base model but worse than other models.

Nevertheless, these observations are not applicable to the tasks from BLiMP. As charts show, on tasks, such as Passive and Principle A, the base model shows better quality than any other models, including the MultiBERT model. At the same time we see that small models encounter difficulties with the acceptability of sentences with transitive verbs and with islands.

The described above leads to the conclusion that bigger models are more successful in language acquisition. Different parameters of model size give different level of improvement. Thus, the most important parameter for language acquisition is hidden size, since it leads to better results for most features. The second best parameter is the number of layers.

Level	Task	Model size
morphology	subject number	768/8
morphology	person	128/2
morphology	passive	512/4
morphology	transitive	512/8
syntax	top constituents	768/8
syntax	tree depth	768/8
syntax	adjunct island	768/8
syntax	principle A	512/4
discourse	discourse coherence	128/2
discourse	Connectors	768/8
discourse	Sentence Position	512/4
discourse	Penn Treebank	128/4

Table 2: The comparison of tasks' acquisition

The results of our experiments with model sizes show that the increase of hidden size has the biggest impact on the quality of models. The number of layers was the second important parameter and improved quality better than the number of attention heads. Our results are similar to the results reported in Wang et al. (2019): they show that larger hidden size tend to improve quality.

The hidden size might be important for smaller models because different layers code different information. For example, Rogers et al. (2020) summarise that the first layers are task-invariant and contain general linguistic information while the latest layers are usually task-specific.

On the contrary, attention heads are usually more detailed, for example, they are known to remember specific syntactic patterns (Htut et al., 2019). Kovaleva et al. (2019) reveal that attention heads learn the same patterns. Therefore, when the resources to encode information are limited, attention heads do not add much new information.

Regarding the hidden size, our results are different from the results in (Wang et al., 2019). While they postulate that number of layers is the most essential parameter, our results show that hidden size is better for performance improvement.

Since on some tasks small models did not reach the level of a base model, we train more models but following the results we achieved in our experiments with model parameters, we limit our experiments to hidden size and number of layers

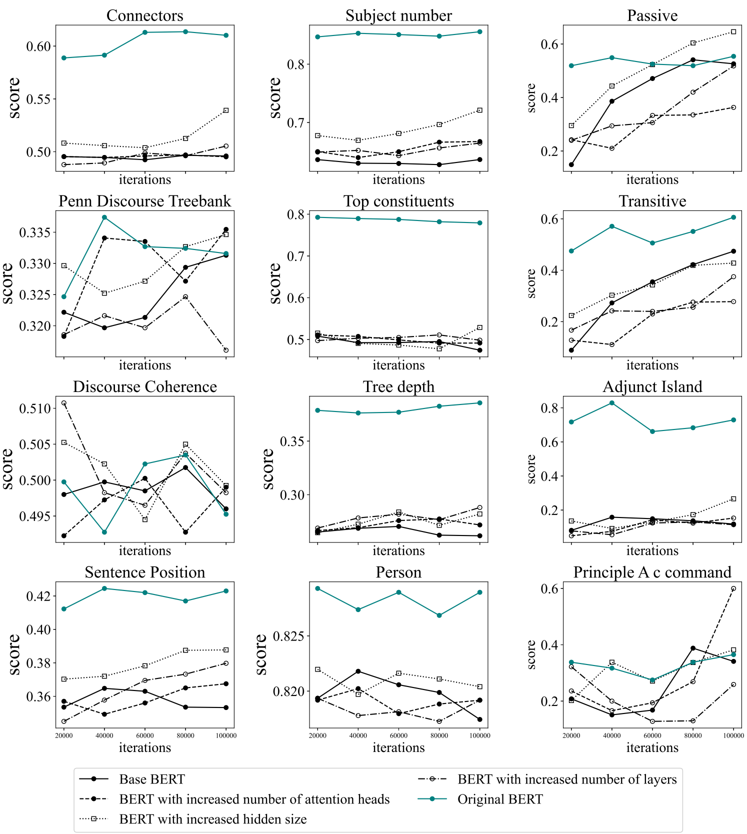

Figure 1: Small models' results on different tasks.

leaving behind the number of attention heads as an insignificant factor.

These experiments summarised in table 2 show that most of the 'morphosyntactic' tasks are acquired by models with hidden size of 768. At the same time, on discourse-based tasks, models with much smaller size show results comparable to the base model.

For most BLiMP tasks the level of the base models is achieved by models of hidden size of 518 with 4 or 8 layers, which is a smaller size than for other 'morphosyntactic' probing tasks.

The results of the experiments prove that increasing hidden size shows better results than increasing number of layers.

Moreover, models with the hidden size of 768 and 8 layers show results close to the model with the same hidden size and 12 layers. Therefore, we conclude that hidden size is the crucial parameter for language acquisition.

5 Conclusion

This works addresses the problem of language acquisition in state-of-the-art models and answers which factors influence the language acquisition process.

To display correlation between language acquisition and different model parameters, we trained four models: one with the minimal hidden size and minimal number of layers and attention heads and three models with one parameter increased and others frozen. These experiments reveal that hidden size appears to be the most essential parameter for language acquisition, whereas attention heads do not significantly increase a model's performance.

Finally, we compared all tasks with the size of a model that shows the quality comparable with the base model used before. The idea behind this comparison is to find any correlation between different language levels and probing measures. As a result, models distinguish discourse from morphology and syntax but there is almost no difference between 'morphological' and 'syntactic' tasks.

References

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2022. Analyzing the mono- and cross-lingual pretraining dynamics of multilingual language models.

Mingda Chen, Zewei Chu, and Kevin Gimpel. 2019. Evaluation benchmarks and learning criteria for discourse-aware sentence representations. In *Proc. of EMNLP*.

Cheng-Han Chiang, Sung-Feng Huang, and Hung-yi Lee. 2020. Pretrained language model embryology: The birth of albert. *arXiv preprint arXiv:2010.02480*.

Noam Chomsky. 1981. Lectures on government and binding (dordrecht: Foris). *Studies in generative grammar*, 9.

Leshem Choshen, Guy Hacohen, Daphna Weinshall, and Omri Abend. 2022. The grammar-learning trajectories of neural language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8281–8297, Dublin, Ireland. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loic Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R Bowman. 2019. Do attention heads in bert track syntactic dependencies? *arXiv preprint arXiv:1911.12246*.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of bert. *arXiv preprint arXiv:1908.08593*.

Ben Krause, Emmanuel Kahembwe, Iain Murray, and Steve Renals. 2019. Dynamic evaluation of transformer language models. *arXiv preprint arXiv:1904.08378*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Leo Z Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A Smith. 2021. Probing across time: What does roberta know and when? *arXiv preprint arXiv:2104.07885*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mitch Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The penn treebank: Annotating predicate argument structure. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Vladislav Mikhailov, Oleg Serikov, and Ekaterina Artemova. 2021. Morph call: Probing morphosyntactic content of multilingual transformers. In *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, pages 97–121, Online. Association for Computational Linguistics.

Allen Nie, Erin Bennett, and Noah Goodman. 2019. Dissent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510.

Tong Niu, Semih Yavuz, Yingbo Zhou, Huan Wang, Nitish Shirish Keskar, and Caiming Xiong. 2020. Unsupervised paraphrase generation via dynamic blocking. *arXiv preprint arXiv:2010.12885*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

John Robert Ross. 1967. Constraints on variables in syntax.

Naomi Saphra and Adam Lopez. 2018. Understanding learning dynamics of language models with svcca. *arXiv preprint arXiv:1811.00225*.

Thibault Sellam, Steve Yadlowsky, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Turc, Jacob Eisenstein, Dipanjan Das, et al. 2021. The multiberts: Bert reproductions for robustness analysis. *arXiv preprint arXiv:2106.16163*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zihan Wang, Stephen Mayhew, Dan Roth, et al. 2019. Cross-lingual ability of multilingual bert: An empirical study. *arXiv preprint arXiv:1912.07840*.

Alex Warstadt and Samuel R Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In *Algebraic structures in natural language*, pages 17–60. CRC Press.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Yian Zhang, Alex Warstadt, Haau-Sing Li, and Samuel R Bowman. 2020. When do you need billions of words of pretraining data? *arXiv preprint arXiv:2011.04946*.

A Examples of Tasks

Task	Sentence examples	Labels
Subject number	*Her employer had escaped with his wife for several afternoons this summer.*	NN
	Your Mackenzie in-laws have sordid reputations few decent families wish to be connected with.	NNS
Person	*So I still can recomend them but prepare pay twice as much as they tell you initially.*	has a person marker
	The service was friendly and fast, but this just does nt make up for the lack - luster product.	does not have a person marker
Tree depth	*We have done everything we can for her .*	11
	Alvin Yeung of Civic Party	3
Top constituents	*Did it belong to the owner of the house ?*	VBD_NP_VP_.
	How long before you leave us again ?	WHNP_SQ_.
Connectors	*He 'd almost forgotten about that man . Sarah had somehow brought him back , just as she had his nightmares .*	but
	I let out a slow , careful breath . Felt tears sting my eyes .	and
Sentence position	*Quneitra Governorate (/ ALA-LC : " Muḥāfazat Al-Qunaytrah ") is one of the fourteen governorates (provinces) of Syria . The governorate had a population of 87,000 at the 2010 estimate . Its area varies , according to different sources , from 685 km ² to 1,861 km ² . It is situated in southern Syria , notable for the location of the Golan Heights . The governorate borders Lebanon , Jordan and Israel .*	1
	The bossom and the part of the xhubleta covered by the apron are made out of crocheted black wool . The bell shape is accentuated in the back part . The xhubleta is an undulating , bell-shaped folk skirt , worn by Albanian women . It usually is hung on the shoulders using two straps . Part of the Albanian traditional clothing it has 13 to 17 strips and 5 pieces of felt .	4
Penn Discourse Treebank	*Solo woodwind players have to be creative,they want to work a lot*	Pragmatic Cause
	The U.S. , along with Britain and Singapore , left the agencyl, its anti-Western ideology , financial corruption and top leadership got out of hand	List
Discourse Coherence	*Within the fan inlet case , there are anti-icing air bosses and probes to sense the inlet pressure and temperature .', 'High speed center of pressure shifts along with fin aeroelasticity were major factors . At the 13th (i.e .', 'the final) compressor stage , air is bled out and used for anti-icing . The amount is controlled by the Pressure Ratio Bleed Control sense signal (PRBC) . The " diffuser case " at the aft end of the compressor houses the 13th stage .*	a text is not coherent
	This experience of digital circuitry and assembly language programming formed the basis of his book " Code : The Hidden Language of Computer Hardware and Software " . Petzold purchased a two-diskette IBM PC in 1984 for $ 5,000 . This debt encouraged him to use the PC to earn some revenue so he wrote an article about ANSI.SYS and the PROMPT command . This was submitted to PC Magazine for which they paid $ 800 . This was the beginning of Petzold 's career as a paid writer . In 1984 , PC Magazine decided to do a review of printers .	a text is coherent

Table 3: Examples of tasks

Task	Acceptable sentence	Unacceptable sentence
Transitive	*The pedestrians question some people.*	*The pedestrians wave some people.*
Passive	*Tracy isn't fired by Jodi's daughter.*	*Tracy isn't muttered by Jodi's daughter.*
Principle A c command	*This lady who is healing Charles wasn't hiding herself.*	*This lady who is healing Charles wasn't hiding himself.*
Adjunct Island	*Who does John leave while alarming Beverly?*	*Who does John leave Beverly while alarming?*

Table 4: BLiMP Minimal pairs examples