Down-Scaling Language Models in the Era of Scale Is All You Need

Anonymous ACL submission

Abstract

Large language models are very resource intensive, both financially and environmentally, and require a huge amount of training data, which is only available to a small number of languages. In this work, we focus on low resource settings. We build language models in two languages trained with different configurations, which are then evaluated on several NLP tasks. Specifically, we analyze three lightweight BERT architectures (with 124M, 51M, and 16M parameters) which are trained with small corpora (125M, 25M, 5M words) for both Basque and Spanish languages. The trained models are evaluated on several tasks, and compared with traditional, non-neural supervised systems. We also present an estimate of resources and CO_2 emissions needed in each approach, which asks for a compromise between raw performance and environmental costs.

1 Introduction

Neural language models (LM) have shown impressive results on many NLP downstream tasks. Nowadays, there is a trend to scale up LMs and build larger and larger ones (Chowdhery et al., 2022), motivated by the fact that large models are known to show emergent abilities that dramatically boost their performance (Wei et al., 2022). However, the environmental cost due to the carbon footprint required to fuel multiple modern GPU hardware is enormous (Strubell et al., 2019). Moreover, building those models requires enormous computational resources, as well as corpora of virtually infinite size, to the extent that, with some notable exceptions[1], only big companies can afford to train such models.

In low resource settings, or when the budget is fixed, one has often to find a compromise on how to spend these resources in an optimal way. Some works have analyzed the performance of LMs wrt. their complexity (in number of parameters) and the size of the corpus on which it is trained (Kaplan et al., 2020; Hoffmann et al., 2022). Those works reveal a positive correlation between these aspects and model efficiency, therefore supporting the current trend followed in LM development. However, they focus on large LMs and training corpora, with the aforementioned problems regarding computational resources and data availability.

It is therefore timely to pay attention to the best strategies under low resource regimes. This paper analyzes the minimum training corpus size and number of parameters needed to build LMs that perform better than non-neural systems. Previous works have already researched on optimal LM parameters within limited resources scenarios, but in most cases with a fixed dataset size (Turc et al., 2019; Raffel et al., 2020) or a fixed model size (Zhang et al., 2021; Hu et al., 2020; Inoue et al., 2021; Martin et al., 2020; Raffel et al., 2020). We complement those works by considering both aspects at the same time, as well as extending the analysis in more than one language.

Our main contributions are the following:

- We address the problem of finding optimal model parameters for a given pre-training corpus in low resource settings.
- We compare the performance of BERT models of different sizes (124M, 51M, 16M), with different corpus sizes (125M, 25M, 5M) for two languages (Basque and Spanish) in MLM and two downstream tasks, NERC and Topic classification.
- Finally, we compare the results of the BERT models with non-neural systems and report the computational costs and CO_2 emissions.

[1] https://huggingface.co/bigscience/bloom

2 Experimental Design

This section describes our experimental setting, which includes 2 languages, 3 corpus sizes, and 3 model sizes, giving a total of 18 different models.

2.1 Language Selection and Corpora

We conduct the experiments in Basque and Spanish, two languages from different language families (although geographically close) that share the Latin alphabet, and fulfil the criteria of having enough monolingual data to train LMs, as well as available evaluation datasets for NLU tasks.

For each language, we created three corpora comprising 125M, 25M and 5M words, respectively. We decided to limit the number of corpora sizes to three in order to control on the number of experiments, and thus the computational resources needed for running them. Preliminary experiments showed a big fall in the results when reducing pre-training data to just 1M words. Since obtaining corpora of 5M words is achievable by most languages that have annotated datasets (Joshi et al., 2020), we set the lower bound at 5M words. The other two corpora sizes are 25 and 125 million, keeping a constant increase rate among them. Corpora in both languages are a mix of 75% news and 25% of text from Wikipedia. We selected the newspaper Berria[2] for Basque, and El Pais[3] for Spanish).

2.2 Models

In a similar fashion to (Turc et al., 2019), we use BERT model in three sizes, dubbed BERT$_{124M}$[4], BERT$_{51M}$ and BERT$_{16M}$, with 12, 8 and 4 layers (L) respectively, also shrinking other hyper-parameters such as hidden dimensions (HH), intermediate layer's dimension (INT) and attention head (H) with the same proportion. Table 1 shows a detailed view of the parameters in each model. We also increase the vocabulary to 50K sub-word tokens, slightly increasing the total number of parameters.

	L	HH	INT	H	Param
BERT$_{124M}$	12	768	3072	12	124M
BERT$_{51M}$	8	512	2048	8	51M
BERT$_{16M}$	4	256	1024	4	16M

Table 1: BERT model sizes used in our experiments. L: num. of layers. HH: hidden dimensions. INT: intermediate layer dimension. H: num. of attention heads.

2.3 Pre-Training Details

We use a cased sub-word vocabulary containing 50K tokens trained with the unigram language model based sub-word segmentation algorithm proposed by Kudo (2018). As mentioned before, we increased the vocabulary sizes of all the models from the original 30K to 50K subword tokens since it seems to be beneficial for agglutinative languages like Basque (Agerri et al., 2020). The vocabularies are learned from each training corpus with a character coverage of 99.95%, to ignore rare characters. Thus, we obtain 3 vocabularies for each language, one for each size of the corpora for pre-training (125M, 25M, 5M), which are shared among LMs of different sizes throughout our experiments.

Input data was duplicated ten times with different masking; we used whole-word masking, where whole words are masked instead of the sub-word units. All models were trained on TPUv3-8 machines using the same set of hyper-parameters in all model sizes: a learning rate $1e^{-4}$, $\beta1$ = 0.9, $\beta2$ = 0.999, L2 weight decay of 0.01, a learning rate warmup of 10K steps, and training the models for a total of 500K steps with a batch size of 256 and a sequence length of 512. See Table 6 for a detailed description of the time spent pre-training each model.

3 Evaluation Settings

3.1 Tasks

We evaluate our models in intrinsic and extrinsic tasks. For the intrinsic evaluation, we tested the models on masked language modeling; for the extrinsic evaluation, we choose two NLU downstream tasks with available datasets in both Spanish and Basque: NERC and topic classification. Table 2 shows the details of each dataset.

Masked Language Modeling (MLM) The first task we selected is masked language modeling, one of default pre-training objectives of BERT, and which is related to the traditional perplexity metric used in language modeling. We report the accuracy of MLM, that is, the number of times that the model correctly guessed the masked token. For this purpose, we created two test datasets from the news domains, from sources not used for the pre-training

[2] https://www.berria.eus

[3] https://elpais.com

[4] This corresponds to BERT$_{base}$ in (Devlin et al., 2019)

Task	Train	Test	Metric
MLM$_{eu}$		1M	acc.
MLM$_{es}$		1M	acc.
NERC$_{eu}$	51,538	35,854	F1
NERC$_{es}$	264,715	51,533	F1
Topic$_{eu}$	8,585	1,854	F1
Topic$_{es}$	9,458	4,000	F1

Table 2: Datasets used to evaluate our models. The size for MLM and NERC is reported in tokens, whereas the size of topic classification datasets is reported in sequences.

MLM$_{eu}$	5M	25M	125M
BERT$_{16M}$	32.08	38.68	41.56
BERT$_{51M}$	32.42	44.29	50.07
BERT$_{124M}$	34.50	43.46	53.19
MLM$_{es}$	**5M**	**25M**	**125M**
BERT$_{16M}$	39.09	49.06	48.31
BERT$_{51M}$	39.24	53.49	59.04
BERT$_{124M}$	42.45	52.58	62.00

Table 3: Accuracies on MLM for Basque and Spanish. Columns correspond to different corpus sizes.

of the models[5].

NERC The second task is Named Entity Recognition and Classification (NERC), a sequence labelling task. For Basque, we selected the in-domain NERC dataset which is part of the benchmark BasqueGLUE (Urbizu et al., 2022). For Spanish, we opted for the Conll2002 dataset (Sang, 2002). We use the F1 score as the performance metric.

Topic Classification The third and last task we selected for evaluation is topic classification, a sequence classification multi-class task. For Basque, we chose the BHTCv2 dataset with 12 classes that was included in BasqueGLUE (Urbizu et al., 2022). And the Spanish counterpart is DVtopic, a dataset[6] we built from news from 8 topics from El Diario Vasco[7]. We use the F1 as the performance metric.

3.2 Systems and Baselines

For the extrinsic evaluation, we fine-tuned each of the 18 models making use of Transformers library (Wolf et al., 2020), training for 10 epochs, with a learning rate of 3^{e-5} and an effective batch size of 32. For each task and language, we report the results as the average of 5 runs.

We compare LMs with traditional, non-neural supervised systems in NERC and topic classification. Regarding NERC, the non-neural system is provided by *ixa-pipe-nerc*[8] (Agerri and Rigau, 2016), which was trained with the same corpora mentioned in Section 3.1. The system consists of language independent local and semi-supervised features based on three types of clustering methods: Brown (Brown et al., 1992), Clark (Clark, 2003) and Word2vec (Mikolov et al., 2013) clustered via K-means. Clusters were trained using the data sources and method described in Agerri and Rigau (2016).

Regarding topic classification, the non-neural system is an SVM classifier trained with documents represented according to a TFIDF model and trained on the corpora mentioned in Section 3.1. Previously, the documents have been lemmatised using Hunspell[9] to reduce the sparseness of the TFIDF vectors.

4 Results

Table 3 shows the results obtained in the MLM task for both Basque and Spanish. As expected, larger models trained with the biggest corpora yield best results, and a correlation exists between model/corpora size and accuracy in both languages. Results also show that overall it is preferable to train a smaller model with more data than a large model using smaller corpora. However, the gain obtained with the smallest BERT$_{16M}$ models as we keep adding training data diminishes, which suggests that performance is reaching a plateau in these models.

The results for the NERC task are shown in Table 4. There is still a clear positive correlation between the evaluation metric and the model and corpora size, but, unlike in the MLM task, BERT$_{51M}$ obtains similar results compared to the largest BERT$_{124M}$ model. Besides, increasing the corpora size is not so helpful, particularly in Spanish, where the training data is large. In any case, even the modest configuration comprising BERT$_{51M}$ and a corpus size of 25M outperforms the non-neural baseline by more than 5 points in Basque and 1.5 points in Spanish. The results for topic classification at Table 5 follow the same trends and present

[5] For Basque we extracted the text from Argia news magazine www.argia.eus; for Spanish, we opted for the newspaper El Mundo www.elmundo.es

[6] Available at: anonymized link

[7] https://www.diariovasco.com

[8] https://github.com/ixa-ehu/ixa-pipe-nerc/

[9] http://hunspell.github.io/

NERC$_{eu}$	5M	25M	125M	NERC$_{es}$	5M	25M	125M
BERT$_{16M}$	63.98±0.4	**74.61**±0.5	**74.40**±0.2	BERT$_{16M}$	76.65±0.4	81.96±0.2	81.63±0.6
BERT$_{51M}$	**74.61**±0.2	**78.95**±0.2	**83.02**±0.3	BERT$_{51M}$	80.83±0.2	**86.11**±0.5	**86.73**±0.2
BERT$_{124M}$	72.92±0.1	**79.50**±0.6	**84.91**±0.2	BERT$_{124M}$	82.21±0.2	**85.80**±0.3	**87.51**±0.2
ixa-pipes	73.95			ixa-pipes	84.16		

Table 4: Results for the 9 models on NERC (F1) for Basque and Spanish. Columns correspond to different corpus sizes. In bold models that outperform the ixa-pipes baseline.

Topic$_{eu}$	5M	25M	125M	Topic$_{es}$	5M	25M	125M
BERT$_{16M}$	**68.26**±0.3	**72.20**±0.5	**72.42**±0.5	BERT$_{16M}$	**84.12**±0.2	**86.48**±0.2	**86.91**±0.2
BERT$_{51M}$	**69.73**±0.6	**72.98**±0.4	**74.61**±0.2	BERT$_{51M}$	**85.20**±0.2	**87.27**±0.2	**88.13**±0.3
BERT$_{124M}$	**71.60**±0.6	**75.19**±0.3	**76.44**±0.3	BERT$_{124M}$	**85.76**±0.4	**87.78**±0.4	**88.88**±0.1
SVM	65.00			SVM	83.00		

Table 5: Results for the 9 models on topic classification (F1) for Basque and Spanish. Columns correspond to different corpus sizes. In bold models that outperform the SVM baseline.

Model	Pre-training (TPUv3-8)	Fine-tuning (RTX3090)	Inference (CPU)
BERT$_{124M}$	∼76h | 98 kgCO$_2$eq	∼91m|17GB |229 gCO$_2$eq	∼651ms
BERT$_{51M}$	∼32h | 41 kgCO$_2$eq	∼39m | 9GB |109 gCO$_2$eq	∼290ms
BERT$_{16M}$	∼10h | 13 kgCO$_2$eq	∼16m | 4GB | 30 gCO$_2$eq	∼166ms
non-neural		∼ 7s (in CPU) |0.09gCO$_2$eq	∼60 ms

Table 6: Computational costs in time and memory for pre-training, fine-tuning and inference and their estimated CO$_2$ emissions. CO$_2$ estimations calculated with *Machine-Learning Impact calculator* (Lacoste et al., 2019). Reported times for fine-tuning correspond to a single run at topic classification in Spanish (the biggest dataset).

even higher gains of the neural systems wrt. the non-neural baseline.

The boost in performance when increasing the model size is larger in downstream tasks than in the MLM intrinsic task, particularly when shifting from the smallest BERT$_{16M}$ to the intermediate BERT$_{51M}$. This indicates that a larger model is better suited for fine-tuning, as the number of trainable parameters is also higher.

Table 6 shows the computational resources and estimated CO$_2$ emissions for each system. Clearly, the non-neural system incurs in the lowest costs, as there is no pre-training needed. Even after pre-training, the CO$_2$ emissions of neural models are order of magnitudes higher, compared to non-neural ones. However, these differences are much smaller at inference time, once the models have been pre-trained and fine-tuned. In any case, the results ask for a compromise between raw performance and computational and environmental costs. It is up to each use case to make the proper decision that balances performance and resource requirements to choose the correct approach.

5 Conclusions

In this paper, we present a study of the performance of small and medium language models using relatively small corpora. We have built up to 18 different combinations of model and corpora size, which have been evaluated on intrinsic and downstream tasks, and have been compared with non-neural supervised systems trained on the same datasets. The experiments show that, overall, the more parameters and training corpus, the better the performance, with significant differences on some tasks. Most LMs outperform non-neural supervised systems, even those based on modest models or pre-trained with reduced corpus. We observe that a BERT$_{51M}$ model and 25M of pre-training data is enough to outperform significantly non-neural systems.

LMs require significant resources, mostly computational but also environmental in the form of CO$_2$ emissions, particularly in the pre-training and fine-tuning phases. All in all, our study shows that in low-resource scenarios certain lightweight configurations of language models are a good alternative to non-neural systems, albeit with a higher computational and environmental cost.

Limitations

Our study is limited to 3 language model sizes and 3 pre-training corpora sizes. Including other model sizes like a BERT-Large or a model between 51M and 16M (where there is a big gap in results), and adding more pre-training corpora sizes (Let's say 625M and 1M words) were out of the scope of this work.

Moreover, we selected two languages for the experimentation. Although they are languages from different language families, including more languages from varied typologies, scripts and characteristics would produce more robust results. The same could be said about including more varied NLU tasks for evaluation.

In addition, we use the default hyper-parameters that are commonly used for BERT-base (BERT$_{124M}$) for the pre-training and fine-tuning of the BERT$_{51M}$ and BERT$_{16M}$ models without any hyper-parameter tuning.

References

Rodrigo Agerri and German Rigau. 2016. Robust multilingual named entity recognition with shallow semi-supervised features. *Artificial Intelligence*, 238(2):63–82.

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. 2020. Give your text representation models some love: the case for basque. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4781–4788.

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Aakanksha et. al Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Alexander Clark. 2003. Combining distributional and morphological information for part of speech induction. In *EACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte De La Clergerie, Djamé Seddah, and Benoît

Sagot. 2020. Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Erik F Sang. 2002. Tjong kim (2002).“introduction to the conll-2002 shared task: Language-independent named entity recognition”. In *COLING-02: The 6th Conference on Natural Language Learning*.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*.

Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, Rodrigo Agerri, and Aitor Soroa. 2022. BasqueGLUE: A Natural Language Understanding Benchmark for Basque. In *Proceedings of the Language Resources and Evaluation Conference*, pages 1603–1612, Marseille, France. European Language Resources Association.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel Bowman. 2021. When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1112–1125.