DeepR3: Reducing, Reusing and Recycling Large Models for Developing Responsible and Green Language Technologies

Aitor Soroa[1], German Rigau[1], Jose M. Alonso-Moral[2], Marcos Garcia[2], Maite Melero[3] and Marta Villegas[3]

[1]HiTZ Basque Center for Language Technology - Ixa NLP Group, University of the Basque Country UPV/EHU
[2]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Spain
[3]Barcelona Supercomputing Center, Barcelona, Spain

Abstract

This paper presents the DeepR3 project, a coordinated project composed of three local projects at the Hitz Centre (University of the Basque Country), CiTIUS (University of Santiago de Compostela) and Barcelona Supercomputing Center, respectively. The main objective of DeepR3 is to research on parameter efficient ways to extend existing pre-trained language models for Spanish, Catalan, Basque, Galician plus English, and adapt them to new domains, genres and languages. In this project, we will apply the newly developed techniques to improve the state of the art on text generation tasks in the mentioned languages, reusing and recycling pre-trained models for machine translation, developing advanced content-based domain applications in sectors such as Meteorology, and developing new benchmarks and datasets for evaluating and assessing progress towards responsible natural language understanding and generation. DeepR3 is funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR.

Keywords

Computational Linguistics, Natural Language Processing, Ethical Guidelines, Trustworthy AI, Sustainable AI

1. Introduction

The Natural Language Processing (NLP) community is currently engaged in a paradigm shift with the production and exploitation of large pre-trained transformer-based language models. Compared to the previous state of the art, the results are so good that systems are claimed to obtain human-level performance when evaluated in difficult language understanding tasks. Some authors call these models "foundation models" to underscore their critically central yet incomplete character. This paradigm shift means that we have only just started to discover the new possibilities and concerns raised by these large pre-trained language models. Despite their impressive capabilities, these models do come with severe drawbacks from a research advancement, environmental, and ethical perspective. These models are extremely costly to train and develop, both financially, due to the cost of hardware and electricity or cloud computing time, and environmentally, due to the carbon footprint required to fuel modern servers with multiple Graphics Processing Unit (GPU) hardware. This also means that only a limited number of organisations with abundant resources in terms of funding, computing capabilities, NLP experts and data can currently afford to develop and deploy such models. A growing concern is that due to unequal access to computing power, only certain firms and elite research groups can access modern Artificial Intelligence (AI) research [1]. We have no clear understanding of how large language models work, when they fail, or what emergent properties they may present, as well as novel ways of exploiting these models efficiently. There are also worrying shortcomings in the text corpora used to train these anglo-centric models, ranging from a lack of representation of less-resource languages such as Catalan, Basque or Galician, to a predominance of harmful stereotypes, and to the inclusion of personal information. To tackle these questions, much critical interdisciplinary collaboration and research are needed.

The DeepR3 project aims to address some of the important concerns that large language models raise from a research, innovation, but specifically from an environmental perspective. Instead of creating very expensive monolingual and multilingual language models from scratch,

SEPLN-CEDI-PD 2024: Seminar of the Spanish Society for Natural Language Processing: Projects and System Demonstrations, June 19-20, 2024, A Coruña, Spain
a.soroa@ehu.eus (A. Soroa); german.rigau@ehu.eus (G. Rigau); josemaria.alonso.moral@usc.es (J. M. Alonso-Moral); marcos.garcia.gonzalez@usc.gal (M. Garcia); maite.melero@bsc.es (M. Melero); marta.villegas@bsc.es (M. Villegas)
https://ixa2.si.ehu.eus/asoroa/ (A. Soroa); https://adimen.si.ehu.es/~rigau/ (G. Rigau); https://citius.gal/es/team/jose-maria-alonso-moral (J. M. Alonso-Moral); https://citius.gal/team/marcos-garcia-gonzalez (M. Garcia); https://www.bsc.es/melero-nogues-maite (M. Melero); https://www.bsc.es/villegas-marta (M. Villegas)
0000-0001-8573-2654 (A. Soroa); 0000-0003-1119-0930 (G. Rigau); 0000-0003-3673-421X (J. M. Alonso-Moral); 0000-0002-6557-0210 (M. Garcia); 0000-0001-9933-3224 (M. Melero); 0000-0003-0711-0029 (M. Villegas)

the aim of the project is to investigate parameter efficient methods to reuse and extend existing pre-trained language models for Spanish, Catalan, Basque, Galician plus English, and adapt them to new domains, genres and languages. The adapted models will be applied to different use cases and tasks such as machine translation, applications in the biomedical domain and text generation from meteorological data. The DeepR3 project may therefore have a direct impact on both the Ecological Transition by avoiding unnecessary waste of energy and the Digital Transition by creating better language models for more languages that can be applied to multiple NLP tasks.

DeepR3 follows the Findable, Accessible, Interoperable and Re-usable (FAIR) principles. First, the results will be used by the project partners which already provide NLP products and services to third party companies. Second, all project results are and will be distributed under open source licenses to many more additional end-users of these technologies (e.g., academia, industry and public administrations). In order to maximize impact and ensure uptake of the methods developed in DeepR3 by the scientific community, the advance in the state of the art is demonstrated by applying a carefully designed evaluation methodology, with results (software, data and extended language models) made publicly available to ensure reproducibility (by hosting datasets, code and data in public repositories such as Huggingface). We are currently involved in the organization of a workshop that is directly aligned with the topic of green language models, collocated with the COLING-LREC2024 conference. A second workshop entitled "Multimodal, interactive and affective eXxplainable AI (XAI)" has been accepted for the European Conference on Artificial Intelligence (ECAI) 2024. It will include a special track on "Building, Reusing, Recycling, and Reducing Resources for Multi-modal XAI".

2. Related Work

Language models that are pre-trained following a self-supervision approach are usually very big. For instance, GPT-3 contains 175 billions of parameters and was trained on 570 gigabytes of text [2], with a cost estimated at more than four million USD[1]. There are even bigger models such as Megatron-530B from NVIDIA, which contains 530 billion trainable parameters which are also the number of activated model parameters per input token [3]. All these models are trained using some variation of the gradient descent method, which requires updating all the parameters of the model at every training iteration. As a consequence, building the models requires weeks and months of processing power. This incurs enormous costs, both financially, due to the cost of hardware and electricity, and environmentally, due to the carbon footprint required to fuel multiple modern GPU hardware. Nowadays, model training and development of large language models are likely to make up a substantial portion of the greenhouse gas emissions attributed to the NLP area. In [4] the authors benchmarked model training and development costs in financial terms and estimated carbon dioxide emissions. While the average human is responsible for an estimated five tons of carbon dioxide per year[2], the authors trained a big neural architecture and estimated that the training process emitted 284 tons of carbon dioxide.

There have been several proposals to address these issues and develop methods that effectively train models without updating every parameter at every training iteration. Many of these techniques have been developed within the framework of distributed or federated learning, where individual workers train the models locally and communicate their changes to a centralized server [5]. In [6] the authors propose a method that chooses a small subset of the model parameters to update, according to the relative importance of the parameter in the final output of the model. By training only on a small fraction of the parameters (as few as 0.5%), they attain similar performance to training all parameters. Certain models can have individual submodels added, removed, or updated while the remainder of the parameters remain fixed. For instance, in [7] authors proposes the mixture-of-experts (MoE) layer, which consists of small feed sub-networks and a trainable gating system that learns how to combine the outputs of the networks for each particular example. At modest training budgets, MoEs can match the performance of dense models using four times less computing effort [8].

In the transfer learning stage, standard fine-tuning updates the whole parameter set of the model, and therefore a language model fine-tuned for a specific task needs to change and store all the parameters of the original model. While not as expensive as the pre-training stage, this parameter inefficiency still incurs high computational and environmental costs, especially in the presence of many downstream tasks. Recently, there has been a growing interest in developing more parameter-efficient fine-tuning. Adapter modules are a light-weight alternative to full model fine-tuning, consisting of only a tiny set of newly introduced parameters at every transformer layer [9]. In [10], the authors propose a new transfer learning method based on merging multiple models into one. For this, they develop a merging method that takes into account the importance of each parameter when computing the average of different models, and find that this form of merging can efficiently transfer knowledge across models fine-tuned from the same pre-trained checkpoint. In [11] authors

[1] https://lambdalabs.com/blog/demystifying-gpt-3/

[2] https://ourworldindata.org/co2-emissions

also show that it is possible to reach similar performances on many downstream-tasks using much smaller language models pre-trained with knowledge distillation, resulting in models that are lighter and faster at inference time, while also requiring a smaller computational training budget [12, 13].

3. Objectives

The main goal of DeepR3 is to obtain substantial scientific and technical contributions in many NLP tasks. In order to achieve this goal, the project proposes:

- Developing efficient methods for extending to new domains, genres and languages, existing language models for the official languages of Spain (Spanish, Catalan, Basque and Galician) plus English.
- Exploring novel ways to pre-train and finetune language models in a parameter-efficient way, therefore lowering the carbon footprint required to train such models.
- Addressing language understanding tasks by text generation.
- Addressing explainability of Deep Learning (DL)-based language models for Natural Language Generation (NLG) tasks.
- Developing new efficient techniques to reuse and recycle pre-trained models for MT, especially for settings with few or non-existing parallel data.
- Applying the newly developed techniques to improve the state of the art in Natural Language Understanding (NLU).
- Developing new benchmarks and datasets for evaluating and assessing progress towards responsible NLU and NLG.
- Developing a number of advanced content-based domain applications for the official languages in Spain (Spanish, Catalan, Basque, and Galician) and English, in multiple sectors and domains (Meteorology, Health, Tourism, Public Administrations, etc).

4. Methodology

The duration of DeepR3 is 24 months. DeepR3 is a coordinated project organized in three sub-projects with shared methodology, objectives and techniques. The general coordination is carried out by the HiTZ-UPV/EHU group. To achieve the objectives explained earlier, the work has been organized in 5 technical Work Packages, plus a WP0 Project Management and a WP6 for Dissemination and Exploitation. Each work package has a responsible leader who coordinates the tasks in the WP, the relationships and dependencies with other WPs, and contributes to checking the overall quality of the project outcomes. Moreover, each task is coordinated by a researcher from the research team.

The technical WPs are briefly introduced below:

WP1 **Methodology and Design.** The purpose of this WP is to define the overall methodology of the project. This includes the specific defining standard protocols, information flow and architectures, adapting and integrating the modules, resources, data structures, data formats, computing facilities and module APIs in the DeepR3 project. The work package will also provide technical coordination and support for the successful achievement of the methodological and technological objectives. In addition, we will consider transversally through the entire project issues related to Responsible AI as well as Ethical, Legal, Social, Economical, and Cultural (ELSEC) perspectives.

WP2 **Data Management.** We target the collection and curation of the corpora and data needed for extending existing monolingual and multilingual language models to new domains, genres and languages, as well as the datasets required to evaluate them. In this WP, we will also address publishing, legal and ethical issues.

WP3 **Adapting compact language models.** We focus on providing efficient ways of adapting pretrained language models to new domains and languages. An important requirement is providing these models in the most compact manner, in terms of model size, to make their use as cost efficient as possible. The goal is to recycle large preexisting models instead of training from scratch and to be environmentally responsible. Attention will be paid to explainability and fairness of the models. The recycled models obtained will be the basis of a new generation of language models for Spanish, Catalan, Basque and Galician.

WP4 **Deployment of language models.** We will deploy the language models obtained in WP3 and adapt them to specific tasks. We will research modular and parameter-efficient finetuning methods as an efficient way to adapt language models to new tasks and languages. The language models developed in WP3 will be the basis to build a new generation of high-level semantic processing tasks in the biomedical domain. Another task that requires new and innovative ways to reuse pre-trained language models is MT in low-resource scenarios. Finally, many other interesting tasks such as summarisation or question answering require generative models.

WP5 Evaluation and Assessment. The objective of this WP is to measure the research progress via objective evaluation metrics and relevant open evaluation campaigns. In addition to standard metrics and comparative analysis for the technologies developed in each WP, we plan to develop datasets and metrics that help to assess the linguistic generalisation capabilities of the language models and resources developed in the project (WP3 and WP4).

5. Current results

This section provides details about work-in-progress which is currently undertaken within the project.

5.1. Scaling Laws in Low-Resource Settings

There have been many works proposing formulas that relate the size of the model, the size of the dataset and the computing budget, the so called "scaling laws". However, these laws are focused on rich languages and a large scale, where the size of the available data is virtually infinite. We have analyzed the effect those variables have on the performance of language models in constrained settings, by building lightweight models trained over a set of small corpora [14]. Our conclusions conclude that the power laws for parameters, data and compute for low-resource settings differ from the optimal scaling laws previously inferred, and data requirements should be higher. Our insights are consistent across all the languages we study, as well as across the MLM and downstream tasks. Furthermore, we experimentally establish when the cost of using a Transformer-based approach is worth taking, instead to favouring other computationally lighter solutions.

5.2. The medical MedMT5

We have built MedMT5, the first open-source text-to-text multilingual model for the medical domain. While there already exist Large Language Models (LLMs) that have been adapted to the medical domain, they have been pre-trained and evaluated with a focus on a single language (English mostly). This is particularly true of text-to-text models, which typically require large amounts of domain-specific pre-training data, often not easily accessible for many languages. MedMT5 has been trained on the largest multilingual corpus for the medical domain in four languages, namely English, French, Italian and Spanish. We also developed two new evaluation benchmarks for all four languages with the aim of facilitating multilingual research in this domain. A comprehensive evaluation shows that MedMT5 outperforms both encoders and similarly sized text-to-text models for the Spanish, French, and Italian benchmarks while being competitive with current state-of-the-art LLMs in English.

5.3. Resources for Linguistic Evaluation of Language Models

Regarding resources for the linguistic evaluation of language models, we have made progress in the area of Targeted Syntactic Evaluation, designing new datasets for Spanish and Galician focused not only on syntactic dependencies but also on the semantic properties of control verbs [15]. Our results show that, although transformer-based models have a high performance in resolving syntactic dependencies, their performance drops dramatically in cases in which syntax and semantics are in interaction. Furthermore, we have created the Galician Parallel Universal Dependencies (PUD) treebank, a new manually annotated corpus for Galician which follows the latest Universal Dependencies guidelines [16]. The fact that the treebank is a parallel corpus translated by professionals makes it an interesting resource for evaluating machine translation models between Galician and other of the 23 languages that have a PUD treebank.

5.4. Reusing, Recycling and Reducing Pre-trained Models for Developing and Evaluating Green Data-to-text Systems: A Use Case with Meteorological Data

We have analyzed empirically how the reuse, recycling, and reduction of pre-trained language models can enhance environmental sustainability in NLP. More precisely, we paid attention to pre-trained models performing sequence-to-sequence text generation in Spanish and Galician with the METEOGALICIA-ES and METEOGALICIA-GL datasets [17]. We developed an experimental pipeline for systematic experimentation, facilitating the definition of baselines, creation of alternative models, text generation, and comprehensive evaluation, which includes assessing values related to energy consumption and the quality of generated text to determine the optimal model. In light of the reported results, we validated the following research hypothesis: "Employing knowledge transfer techniques enables the creation of low-cost language models that yield results equivalent to or superior to those produced by baseline models with a higher computational cost."

5.5. An Empirical Study on the Number of Items in Human Evaluation of Automatically Generated Texts

We have analyzed empirically the effect of the number of items, i.e., texts to be evaluated, for the sake of reproducibility in human evaluation of NLG systems. In light of the reported results, we validated the following research hypothesis: "well-known resampling statistical methods can contribute to getting significant results even with a small number of items to be evaluated by each evaluator." [18]

5.6. Enriching Interactive Explanations with Fuzzy Temporal Constraint Networks

We have designed, implemented and validated a new model for fuzzy temporal reasoning to overcome some inconsistencies detected in pre-trained language models [19]. As a proof of concept, the new model is integrated with TimeVersa, a conversational agent carefully designed for acting as a virtual assistant for tourists. It handles imprecise temporal constraints when providing users with multilingual recommendations, and related explanations, which are endowed with a good balance between naturalness and fidelity. Taking as starting point a knowledge graph that provides an intuitive representation of the entities and relations in the application domain, the temporal information is mapped onto a fuzzy temporal constraint network. This way we represent imprecise temporal information and are able to check consistency in conversations.

6. Further work

We plan to work on the main areas of the project with main attention to extending and reusing existing language models and adapting them to new new domains, genres and languages, and with a special focus on the languages spoken in the Iberian peninsula. We will continue the work on the generation of data-to-text systems, and extend it to new geographical areas such as the Basque Country. We also plan to develop machine translation systems for Iberian languages and English by reusing existing pre-trained models in these languages. Finally, we will continue developing benchmarks and datasets for evaluating and assessing progress towards responsible NLP, regarding both NLU and NLG tasks.

Acknowledgments

This publication is supported by Grants TED2021-130295B-C31, TED2021-130295B-C32, and TED2021-130295B-C33 funded by MCIN/AEI/10.13039/501100011033 and by the "European Union NextGenerationEU/PRTR".

References

[1] N. Ahmed, M. Wahed, The de-democratization of ai: Deep learning and the compute divide in artificial intelligence research, 2020. arXiv:2010.15581.

[2] T. Brown, B. Mann, N. Ryder, e. a. Subbiah, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, 2020, pp. 1877–1901.

[3] S. Smith, M. Patwary, B. Norick, P. LeGresley, S. Rajbhandari, J. Casper, Z. Liu, S. Prabhumoye, G. Zerveas, V. Korthikanti, et al., Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model, arXiv preprint arXiv:2201.11990 (2022).

[4] E. Strubell, A. Ganesh, A. McCallum, Energy and policy considerations for deep learning in NLP, in: Proceedings of the ACL, Association for Computational Linguistics, Florence, Italy, 2019, pp. 3645–3650.

[5] A. F. Aji, K. Heafield, Sparse communication for distributed gradient descent, in: M. Palmer, R. Hwa, S. Riedel (Eds.), Proceedings of EMNLO, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 440–445.

[6] Y.-L. Sung, V. Nair, C. A. Raffel, Training neural networks with fixed sparse masks, in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J. W. Vaughan (Eds.), Advances in Neural Information Processing Systems, volume 34, Curran Associates, Inc., 2021, pp. 24193–24205.

[7] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, J. Dean, Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, in: Proceedings of ICLR, 2017.

[8] M. Artetxe, S. Bhosale, N. Goyal, T. Mihaylov, M. e. a. Ott, Efficient large scale language modeling with mixtures of experts, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 11699–11732.

[9] J. Pfeiffer, I. Vulić, I. Gurevych, S. Ruder, MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 7654–7673.

[10] M. Matena, C. Raffel, Merging models with fisher-weighted averaging, 2022. arXiv:2111.09832.
[11] V. Sanh, A. Webson, C. Raffel, S. Bach, L. Sutawika, e. a. Zaid Alyafeai, Multitask prompted training enables zero-shot task generalization, in: Proceedings of ICRL, 2022.
[12] C. Bucilâ, R. Caruana, A. Niculescu-Mizil, Model compression, in: Proceedings of KDD, KDD '06, Association for Computing Machinery, New York, NY, USA, 2006, p. 535–541.
[13] G. E. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network., CoRR abs/1503.02531 (2015).
[14] G. Urbizu, I. S. Vicente, X. Saralegi, R. Agerri, A. Soroa, Scaling laws for bert in low-resource settings, in: Findings of the ACL, 2023.
[15] I. de Dios-Flores, J. Garcia Amboage, M. Garcia, Dependency resolution at the syntax-semantics interface: psycholinguistic and computational insights on control dependencies, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of ACL, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 203–222.
[16] X. Sánchez-Rodríguez, A. Sarymsakova, L. Castro, M. Garcia, Increasing manually annotated resources for Galician: the Parallel Universal Dependencies Treebank, in: Proceedings of PROPOR, volume 1, Association for Computational Linguistics, 2024, pp. 694–699.
[17] J. González-Corbelle, J. Alonso-Moral, A. B. Diz, J. Taboada, Dealing with hallucination and omission in neural Natural Language Generation: A use case on meteorology, in: Proceedings of NLG, ACL, 2022, pp. 121–130. Https://aclanthology.org/2022.inlg-main.10.
[18] J. Gonzalez-Corbelle, J. M. Alonso-Moral, R. M. Crujeiras, A. Bugarín-Diz, An empirical study on the number of items in human evaluation of automatically generated texts, Procesamiento del Lenguaje Natural (2024).
[19] M. Canabal-Juanatey, J. M. Alonso-Moral, A. Catala, A. Bugarín-Diz, Enriching interactive explanations with fuzzy temporal constraint networks, International Journal of Approximate Reasoning (2024) 109128.