Empirical Study on the Energy Efficiency of Transfer Learning Techniques for Text-to-Text Generation

Ainhoa Vivel-Couso

Second year PhD student at Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Department of Electronics and Computing, University of Santiago de Compostela, Spain
Supervised by Jose Mª Alonso Moral and Alberto Bugarín-Diz
ORCID (Ainhoa Vivel-Couso): https://orcid.org/0000-0002-5860-4849

Abstract. This PhD project focuses on promoting environmental sustainability in Natural Language Processing (NLP) by emphasising the reuse, recycling, and reduction of pre-trained language models. The main objectives are (i) to improve existing multilingual models through efficient adaptation methods for new domains, genres, and languages; (ii) to investigate approaches for adapting language models that minimize their carbon footprint; and (iii) to develop a systematic experimentation framework that includes baseline definition, alternative model creation, text generation, and comprehensive evaluation to balance energy consumption with the quality of generated text and to identify optimal models. This work underscores the importance of environmental responsibility in NLP and includes a case study on generating textual meteorological forecasts in Spanish, Galician, and Basque.

1 Research problem and motivation

There exists an increasing trend in the scientific community in favour of the idea that training Artificial Intelligence (AI) models should be considered within the framework of the 17 Sustainable Development Goals [26] due to its social, economic, and environmental implications, being the high demand of significant computational power and energy resources one of the key elements of this impact [4].

Training large AI models can be expensive due to the requirement for powerful hardware, specialised processors such as Graphics Processor Units (GPUs), Tensor Processing Units (TPUs), and/or cloud computing resources [23]. Techniques like transfer learning aim to reduce the need for extensive training by leveraging pre-trained models on similar tasks. Ongoing research aims to develop techniques that balance the need for accurate models with environmental and economic considerations. Green AI initiatives specifically target the creation of sustainable and eco-friendly AI technologies, addressing the environmental impact of AI research and applications [22].

The primary goal of this work is to explore knowledge transfer methods to reduce the environmental impact of AI model training. To this end, a baseline model trained under traditional standards will be established. Its performance will be compared to that of alternative models created using various knowledge transfer techniques and less resource-intensive training methods. The study focuses on sequence-to-sequence text generation models in Spanish, Galician, and Basque, specifically adapting them to generate meteorological narratives.

2 Related work

In spite of its importance, there are only a few recent publications in the literature where the environmental impact of large language models (LLMs) is analyzed [12, 21, 28]. The proposal in our research presents a new approach to the problem, since it addresses the comparison of knowledge transfer methods to reduce environmental impact.

2.1 Natural Language Generation

Within the myriad of existing Natural Language (NLG) technologies and models [7], we will focus on the use of pre-trained models [6, 27] which can be fine-tuned [5] for specific tasks. Pre-trained models have become popular in various domains due to their ability to capture and transfer knowledge from diverse datasets, improving efficiency and performance for specific applications [24]. In this work, we will use Sequence-to-Sequence (seq2seq) language models [16], since this type of pre-trained Text-to-Text (T2T) systems have been successfully reused to generate weather descriptions like the ones we will consider as use case.

2.2 Transfer Learning

Transfer learning (TL) [25] is a machine learning (ML) technique where a model trained on one task is adapted to another related task. Thus, TL leverages the knowledge gained from solving a different but related task. This approach is particularly useful when we have a limited amount of labeled data for the target task, and it is especially important to reduce the energy cost and carbon footprint.

TL is valuable in deep learning (DL), where models have many layers and parameters. Popular pre-trained models, such as those based on convolutional neural networks (CNNs) for image-related tasks or models like BERT for NLP, have shown success in TL scenarios [15].

Fine-tuning (FT) is the process of making small adjustments to achieve the desired output or performance [5]. In the context of DL, FT involves the use of weights of a trained neural network to program another DL algorithm from the same domain. Below, we go deeper with those methods which are the most pertinent for the scope of this work:

- **Zero-shot Learning** (ZSL) is a setup in DL where, at test time, a learner observes samples from classes which were not observed during training [11, 14]. ZSL showcases the generalisation and adaptability of pre-trained language models to a wide array of NLP tasks. Nevertheless, even if ZSL offers significant advantages, it may have limitations in cases where the task description is ambiguous or the model's pre-trained knowledge does not align well with the target task.
- **Few-shot Learning** (FSL) is an ML method ready to exploit a training dataset with limited information [2, 10]. FSL is an alternative approach to FT with very limited labelled data.
- **Adapter-based Learning** (ADT). The pre-trained model will not be retrained. Instead, we will introduce an adapter module [18]. Parameter-Efficient Fine Tuning (PEFT) methods involve freezing the pre-trained model parameters and introducing a small number of trainable parameters, known as adapters, on top of it. These adapters are trained to capture task-specific information. We used Low Rank Adapters [8] (LoRA), a technique that accelerates training of large models while consuming less memory.

2.3 Metrics for Text Evaluation

There are numerous metrics for evaluating the quality of language models [3, 13]. The choice of metrics depends on the specific task or goal since different metrics capture various aspects of text quality. Text evaluation metrics can be categorised into human evaluation [20], which entails the application of human judgment to subjectively assess various aspects of the text quality (e.g., fluency, coherence, relevance, or overall quality), and automatic evaluation metrics, which employ computational methods to assess the quality of generated text versus reference or target text.

On the one hand, Perplexity [1] is not commonly used to evaluate seq2seq models like mT5 (Multilingual Translation Transformer) [29], where the task involves transforming an input sequence into an output sequence. On the other hand, metrics such as BLEU (Bilingual Evaluation Understudy) [17] or ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [9] are the most used despite the ongoing debate about their adequacy for NLG tasks [19].

3 Contributions

As I am currently in the second year of this PhD, I am still in the early stages of my research. Nevertheless, several contributions have already been achieved, laying the foundation for subsequent work.

Our current project is a continuation of my master's thesis. This research is being developed in collaboration with HiTZ[1], the Basque Centre for Language Technologies, which brings expertise in multilingual NLP, speech processing, and language resources, which complements my focus on sustainable AI. Through this collaboration, we aim to create more efficient and eco-friendly language models. Together, we plan to develop robust adaptation methods and evaluation frameworks that address both performance and environmental impact.

We are now working on developing and validating the necessary tools to pave the way for Responsible NLP technology. A central contribution of this work is the definition of a dynamic experimental cycle that iteratively covers research, design, development, validation and optimisation. This framework has already guided the construction of a baseline and subsequent transfer learning models. To ensure flexibility, it has been designed to work with different models and datasets while remaining easy to configure. The framework created implements the experimental process comprising the following stages:

- Standby Power Measurement. This involves quantifying the energy consumption of the GPUs when no processes are running. This measurement is expressed in watts-hour.
- Data Preprocessing. It is done to generate the datasets used for defining the Baseline and training the models.
- Baseline Definition. The training involves establishing a reference model to serve as a baseline. This model will undergo traditional and resource-intensive training.
- Knowledge Transfer. The generation process involves creating alternative language models through less resource-intensive training, employing diverse knowledge transfer techniques.
- Automatic Text Generation. Automatically generating narratives from test data for each trained model, including the Baseline.
- Evaluation. Evaluating texts generated by all models, including the Baseline, using automatic metrics.

We conducted preliminary experiments to evaluate the framework. A baseline and several alternative models were defined, and the results are presented in Table 1.

TL technique	TData	Epochs	TTime	TLoss	EEC
ZSL (mT5-base)	-	-	0	-	0
FT (Baseline)	ES	300	12025	0.118	1690
FSL	ES	5	84	1.973	10
FSL	ES	25	396	0.938	56
FT	ES	5	200	1.699	28
FT	ES	25	994	0.843	139
FSL	GL	5	99	4.365	12
FSL	GL	25	473	1.169	66
FT	GL	5	221	1.813	31
FT	GL	25	1103	0.972	154
FSL	ES-GL	5	175	2.227	24
FSL	ES-GL	25	866	0.865	121
FT	ES-GL	5	404	1.578	56
FT	ES-GL	25	2013	0.744	285

Table 1. Training Time in seconds (TTime), Training loss (TLoss), and Estimated Energy Consumption (EEC) in watts-hour (Wh) of each model resulting from training mT5-base with different TL techniques.

4 Conclusions and Future Work

For now, we have focused on meteorological data. Our observations indicate that: (i) refined language models for Spanish and Galician demonstrate adaptability to a new domain through efficient TL techniques; (ii) exploring strategies to minimize the carbon footprint in model adaptation supports a commitment to environmentally conscious practices in NLP; and (iii) the pipeline developed for systematic experimentation has proven effective in defining baselines, creating alternative models, generating text, and conducting a thorough evaluation, including metrics for energy consumption and text quality. Overall, applying knowledge TL techniques enables the creation of low-cost language models that equal or surpass the performance of the baseline model.

For future work, we plan to conduct a more detailed comparison that incorporates a wider range of automatic evaluation metrics, allowing for a comprehensive assessment of model output. We also aim to perform human evaluation, providing a more nuanced perspective on model effectiveness. In addition, we intend to foster ongoing improvement and innovation toward refining language models for sustainable and effective NLP applications.

[1] https://www.hitz.eus/

References

[1] Y. Bengio, R. Ducharme, and P. Vincent. A neural probabilistic language model. *Advances in neural information processing systems*, 13, 2000.

[2] T. Brown et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[3] Y. Chang et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 2023. https://doi.org/10.1145/3641289.

[4] Z. Chen, M. Wu, A. Chan, X. Li, and Y.-S. Ong. Survey on AI sustainability: Emerging trends on learning algorithms and research challenges. *IEEE Computational Intelligence Magazine*, 18(2):60–77, 2023. https://doi.org/10.1109/MCI.2023.3245733.

[5] J. Dodge et al. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv:2002.06305*, 2020. https://doi.org/10.48550/arXiv.2002.06305.

[6] S. Edunov et al. Pre-trained language model representations for language generation. 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 4052–4059, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1409. URL https://aclanthology.org/N19-1409.

[7] A. Gatt and E. Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170, 2018.

[8] E. Hu et al. LoRA: Low-rank adaptation of large language models. *arXiv:2106.09685*, 2021. https://doi.org/10.48550/arXiv.2106.09685.

[9] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. https://aclanthology.org/W04-1013.

[10] X. Lin et al. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates, 2022. ACL. doi: 10.18653/v1/2022.emnlp-main.616. URL https://aclanthology.org/2022.emnlp-main.616. https://doi.org/10.18653/v1/2022.emnlp-main.616.

[11] J. Lu et al. What makes pre-trained language models better zero-shot learners? In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2288–2303, Toronto, Canada, 2023. ACL. doi: 10.18653/v1/2023.acl-long.128. URL https://aclanthology.org/2023.acl-long.128. https://doi.org/10.18653/v1/2023.acl-long.128.

[12] A. Luccioni, S. Viguier, and A.-L. Ligozat. Estimating the carbon footprint of bloom, a 176b parameter language model. *Journal of Machine Learning Research*, 24(253):1–15, 2023.

[13] G. Melis, C. Dyer, and P. Blunsom. On the state of the art of evaluation in neural language models. *arXiv:1707.05589*, 2017.

[14] Y. Meng, J. Huang, Y. Zhang, and J. Han. Generating training data with language models: Towards zero-shot language understanding. *Advances in Neural Information Processing Systems*, 35:462–477, 2022.

[15] M. Mozafari, R. Farahbakhsh, and N. Crespi. A BERT-based transfer learning approach for hate speech detection in online social media. In *Proceedings of the Eighth International Conference on Complex Networks and Their Applications*, pages 928–940. Springer, 2020.

[16] G. Neubig. Neural machine translation and sequence-to-sequence models: A tutorial. *arXiv:1703.01619*, 2017. https://doi.org/10.48550/arXiv.1703.01619.

[17] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318. ACL, 2002. https://doi.org/10.3115/1073083.1073135.

[18] J. Pfeiffer et al. AdapterHub: A framework for adapting transformers. In Q. Liu and D. Schlangen, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54. ACL, 2020. doi: 10.18653/v1/2020.emnlp-demos.7. URL https://aclanthology.org/2020.emnlp-demos.7. https://doi.org/10.18653/v1/2020.emnlp-demos.7.

[19] E. Reiter. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401, 2018. doi: 10.1162/coli_a_00322. URL https://aclanthology.org/J18-3002.

[20] E. Reiter, R. Robertson, and L. Osman. Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, 144(1):41–58, 2003. ISSN 0004-3702. doi: https://doi.org/10.1016/S0004-3702(02)00370-3. URL https://www.sciencedirect.com/science/article/pii/S0004370202003703. https://doi.org/10.1016/S0004-3702(02)00370-3.

[21] M. Rillig et al. Risks and benefits of large language models for the environment. *Environmental Science & Technology*, 57(9):3464–3466, 2023. https://doi.org/10.1021/acs.est.3c01106.

[22] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni. Green AI. *Communications of the ACM*, 63(12):54–63, 2020.

[23] E. Strubell, A. Ganesh, and A. McCallum. Energy and policy considerations for deep learning in NLP. In A. Korhonen, D. Traum, and L. Marquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650. "ACL", 2019. doi: 10.18653/v1/P19-1355. URL https://aclanthology.org/P19-1355. https://doi.org/10.18653/v1/P19-1355.

[24] R. Tinn et al. Fine-tuning large neural language models for biomedical natural language processing. *Patterns*, 4(4), 2023. https://doi.org/10.1016/j.patter.2023.100729.

[25] L. Torrey and J. Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI Global, 2010.

[26] United Nations. Transforming our world: The 2030 agenda for sustainable development. https://sdgs.un.org/2030agenda, 2015.

[27] H. Wang et al. Pre-trained language models and their applications. *Engineering*, 2022. https://doi.org/10.1016/j.eng.2022.04.024.

[28] L. Weidinger et al. Taxonomy of risks posed by language models. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229, 2022.

[29] L. Xue et al. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, 2021. ACL. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41.