The Energy and Carbon Footprint of Training End-to-End Speech Recognizers

Titouan Parcollet, Mirco Ravanelli

The Energy and Carbon Footprint of Training End-to-End Speech Recognizers

Titouan Parcollet[1,2]*, Mirco Ravanelli*[3]

[1]LIA, Avignon Université, France
[2]University of Cambridge, United Kingdom
[3]Mila, Université de Montréal, Canada
titouan.parcollet@univ-avignon.fr

Abstract

Deep learning contributes to reaching higher levels of artificial intelligence. Due to its pervasive adoption, however, growing concerns on the environmental impact of this technology have been raised. In particular, the energy consumed at training and inference time by modern neural networks is far from being negligible and will increase even further due to the deployment of ever larger models.

This work investigates for the first time the carbon cost of end-to-end automatic speech recognition (ASR). First, it quantifies the amount of CO_2 emitted while training state-of-the-art (SOTA) ASR systems on a university-scale cluster. Then, it shows that a tiny performance improvement comes at an extremely high carbon cost. For instance, the conducted experiments reveal that a SOTA Transformer emits 50% of its total training released CO_2 solely to achieve a final decrease of 0.3 of the word error rate. With this study, we hope to raise awareness on this crucial topic and we provide guidelines, insights, and estimates enabling researchers to better assess the environmental impact of training speech technologies.

Index Terms: carbon footprint, end-to-end speech recognition.

1. Introduction

Atmospheric concentrations of carbon dioxide, methane, and nitrous oxide are at alarming levels [1]. Together with other anthropogenic factors, they most likely led us to a climate crisis involving drastic changes in our ecosystem [2, 3]. Alongside a growing interest in using artificial intelligence (AI) to tackle climate change [4], numerous concerns involving the carbon footprint of deep learning (DL) started to emerge [5–9]. Its worldwide adoption, the deployment of larger neural models, the increase in available data, the potential inefficiency of considered computational resources, and the slow uptake of renewable energies in numerous countries are all concomitant factors that will likely result in important environmental costs [6, 7, 10].

Deep learning architectures are commonly trained for dozens, hundreds, or sometimes thousands of hours on specialized hardware accelerators in data centers, that are known to be extremely energy-consuming [11]. According to [12], such a trend is not going to end soon as the demand for AI computing has grown by more than $300,000\times$ from 2012 to 2018. In this context, recent studies have been conducted to highlight the carbon emissions of very large-scale experiments [6, 7, 10]. Nevertheless, such studies are often disconnected from the experiments routinely performed by researchers. Extremely large models such as GPT-3 [13] or switch Transformers [14] that contain billions or trillions of parameters are not related to the environments commonly employed within the community.

Table 1: *Estimated CO_2 from training a single French (CommonVoice) or English (LibriSpeech) state-of-the-art end-to-end speech recognizer on Nvidia Tesla V100 GPUs.*

Models	**CO_2 kg**
French with training in France (avg)	2.8
French with training in Australia (avg)	34.7
English with training in France (avg)	9.4
English with training in Australia (avg)	118
Examples	
Driving a car for 100 km (EU avg)	12.2
CO_2 per capita in 1 year (EU avg)	6700

Furthermore, the precise carbon cost of popular deep learning applications, such as automatic speech recognition (ASR) remains largely unexplored. Such in-domain carbon estimates are essential to enable researchers to quantify the carbon footprint of their daily experiments. In a standard development process, an optimal ASR model often comes from numerous iterative or concurrent training runs due to the need for hyperparameter tuning, support of multiple languages, or simply to be competitive with the state-of-the-art performance. Hence, it becomes crucial to quantify the CO_2 emitted while training these models and assess the trade-off between their carbon footprint and performance.

This work aims at assessing and analyzing the carbon footprint induced by training medium-scale ASR models that are prevalent within the recent literature. Example estimates are given in Table 1. More precisely, we focus on SOTA end-to-end (E2E) systems, that rely on popular architectures including Transformers [15], deep convolutional neural networks [16] or recurrent sequence transducers [17]. The contributions of this work are three-fold:

1. Provide precise CO_2 emissions estimates of three SOTA E2E ASR systems on two well-known tasks (LibriSpeech and CommonVoice French). These estimates are obtained by considering the energy efficiency of a medium-scale university cluster equipped with popular GPUs (Nvidia Tesla V100 and Nvidia RTX 2080 Ti).
2. Analyze the trade-off between CO_2 emissions and word error rate performance during training.
3. Release all pre-trained models to the community through SpeechBrain[1] and HuggingFace [18] to facilitate replicability and encourage further experiments.

[1]https://speechbrain.github.io/

The conducted experiments suggest that CO_2 emissions of SOTA end-to-end speech recognizers are far from being negligible, especially considering that such models are often trained multiple times within the speech community. As expected, the obtained CO_2 estimates vary significantly from country to country. For instance, training a Transformer ASR on LibriSpeech in France emits 9.7 kg of CO_2 compared to 118 kg in Australia.

Furthermore, the experiments also highlight a non-linear relationship between CO_2 emissions and the word error rate (WER) obtained with the considered model. The employed Transformer ASR produces half of its total emissions to solely reduce its WER by 0.3. Such a difference in WER, however, is necessary to achieve SOTA performance. This naturally leads to the following question: *is a tiny performance improvement worth doubling the carbon footprint?*

2. Quantifying CO_2 emissions

The process of estimating the amount of CO_2 released during the training of deep learning models can be summarized in two steps. First, it is necessary to compute the total energy consumed by the hardware and the infrastructure (Section 2.1). Then, this amount is converted to CO_2 emissions based on the geographical location of the resources (Section 2.2).

2.1. Energy Consumption

At training time, the vast majority of the energy consumption comes from GPUs and CPUs [6]. A first approach could be to estimate the consumption based on the hardware power specification assuming a full GPU/CPU usage. However, such a scenario is not realistic as the hardware is usually partially used during training. To alleviate this issue, we repeatedly query system information to obtain the instantaneous consumption of the devices as proposed in the *CarbonTracker* tool [7].

The second largest source of energy used in datacenters comes from cooling (*e.g* up to 40% [19]). Estimating this quantity properly is challenging as it strongly depends on the datacenter infrastructure. A common approach is to consider the Power Usage Effectiveness (PUE). The PUE is the ratio between the total amount of energy used by the facility ($p_{facility}$) and the energy delivered to compute nodes ($p_{compute}$):

$$\text{PUE} = \frac{p_{facility}}{p_{compute}}. \tag{1}$$

As reported in the *2020 Data Center Industry Survey Results* [20], the world average PUE for the year 2020 was 1.59. As expected, the PUE strongly depends on the actual computing infrastructure. For instance, *Google* declares a comprehensive trailing twelve-month PUE ratio of 1.11 [21] compared to 1.2 and 1.125 for *Amazon* [22] and *Microsoft* [23], respectively. Unfortunately, such low PUE are not representatives of the academic field that mainly relies on smaller servers or clusters. In this work, we use the PUE ratio that we computed on the Avignon University (AU) cluster ($PUE_{AU} = 1.55$). This cluster is a medium-size computing infrastructure composed of four computational bays, which we think is representative of university clusters. It is important to note that other hardware components may be responsible for energy consumption, such as RAM or HDD. According to [24], one may expect a variation of around 10% while considering these parameters. However, they are also highly dependent on the infrastructure and therefore discarded in our analysis. Finally, the amount of energy consumed during a training of duration d is given by:

$$e_{total} = \text{PUE} \int_0^d (p_g + p_c), \tag{2}$$

with p_g and p_c the instantaneous power consumption of all GPU and CPU devices.

2.2. CO_2 Conversion

The CO_2 conversion rate c_{rate} is defined as the CO_2 (in grams) emitted for each kWh of energy consumed. This factor largely varies across countries and is strongly linked to the production nature of electricity (*e.g* nuclear *vs* gaz) [10, 24]. Fortunately, it is common for governments to release this coefficient in their official statistics on environmental policies assessment [25, 26]. For instance, France has a C_{rate} of 52 *gCO_2/kWh* [25] compared to 417 *gCO_2/kWh* for the USA [26] or 656*g CO_2/kWh* in Australia [27]. The carbon rate is then integrated with Eq. 2 to compute the total amount of CO_2 emitted as:

$$T_{carbon} = c_{rate} e_{total}. \tag{3}$$

Carbon emissions may be compensated by carbon offsetting or with the purchases of Renewable Energy Credits (RECs, in the USA) or Tradable Green Certificates (TGCs, in the EU). Carbon offsetting allows polluting actions to be compensated directly via different investments in *environment-friendly* projects, such as renewable energies or massive tree planting [28]. RECs and TGCs [29], on the other hand, guarantee that a specific volume of electricity of their owner is generated from renewable energy sources. More precisely, RECs and TGCs purchasing aims to create more renewable energy in the long-term according to the *additionality* principle, by creating an increased demand for these energy sources. In the specific context of CO_2 estimates, RECs and TGCs enable institutions and companies to voluntarily decrease their individual C_{rate}. However, in our analysis, carbon rates are obtained at country level and do not integrate industry level carbon offsetting schemes or RECs.

3. E2E ASR Models and Protocol

To provide impactful estimates, this study analyzes popular setups that are common within the speech community. First, it is important to note that all considered ASR models were trained prior to the estimation process. Therefore, and to limit the carbon footprint of our analysis, measurements are performed across 3 training epochs mirroring the original hyperparameters of the considered model to alleviate any variation in the results. Obtained values are then averaged and multiplied by the number of epochs needed by the original ASR system to reach the final performance. The latter setup enables us to quickly compute precise estimates from different hardware and experimental environments.

Hardware and estimation parameters. We consider two GPUs commonly used by deep learning researchers and engineers: the Nvidia Tesla V100 32GB (*i.e* high-end) and Nvidia RTX 2080 Ti (*i.e* mid-tier). Two Intel Xeon Silver 4210R and two Intel Xeon E5-2698 v4 CPUs are attached to Tesla and RTX GPUs respectively. The PUE of 1.55 computed for the Avignon Université cluster is used across all experiments combined with French and Australia conversion rates (Section 2.2). Power draw measurements are obtained every five seconds and then averaged within *CarbonTracker* [7].

Table 2: *CO_2 and energy consumption estimates for popular E2E ASR models trained with computational resources located in France or in Australia (Au.). The Word Error Rates (WER) on CommonVoice (CV) FR and LibriSpeech (LS) are obtained on the "test" and "test-clean" sets of CV and LS respectively. The (x,y) given with model name indicates the number of GPUs used for (CV,LS).*

	CommonVoice					LibriSpeech				
Tesla V100	**kWh per epoch**	**Epochs**	**CO_2 France (kg)**	**CO_2 Au. (kg)**	**WER %**	**kWh per epoch**	**Epochs**	**CO_2 France (kg)**	**CO_2 Au. (kg)**	**WER %**
CRDNN (1,1)	2.11	25	2.77	34.66	17.70	3.78	25	4.92	62.04	2.90
Transformer (1,1)	0.92	40	1.92	24.20	20.57	1.49	121	9.38	118.4	2.55
RNN-T (1,3)	2.00	30	3.12	39.35	20.18	6.58	30	10.26	129.5	5.23
RTX 2080 Ti										
CRDNN (3,3)	5.28	25	6.87	86.63	17.70	5.40	25	7.01	88.45	2.90
Transformer (3,8)	2.98	40	6.19	78.13	20.57	1.72	121	10.87	137.1	2.55
RNN-T (3,6)	4.37	30	6.83	86.16	20.18	8.37	30	13.06	164.7	5.23

Speech recognition datasets. Two datasets with different complexity and size are used as benchmarks. First, we trained our models on the 960 hours of LibriSpeech and evaluated them on the official *"dev-clean"* and *"test-clean"* subsets [30]. Then, we considered the French corpus of CommonVoice [31] (version 6.1) composed of 438 hours of read speech for training. CommonVoice is challenging as it contains realistic recording conditions that vary significantly across speakers. The official validation and test splits are used for evaluation.

End-to-end ASR models. We considered three very popular E2E ASR systems able to reach state-of-the-art performance:

1. *CTC-Attention based Transformers* [15]: a model based on a CNN-Transformer encoder with a Transformer decoder jointly trained with the CTC loss.
2. *CTC-Attention based CRDNN* [32]: an encoder-decoder ASR system. The encoder is composed of three distinct parts: a VGG-like features extractor, a bidirectional LSTM, and a deep dense neural network. This is combined with a location-aware attentive recurrent (GRU) decoder jointly trained with the CTC loss.
3. *RNN-Transducers* [33]: The encoder is a CRDNN while the prediction network is a GRU network. The joint nonlinear layer takes as input the concatenation of the output from both networks. The entire system is trained following the transducer loss.

LibriSpeech experiments rely on an external language model (LM) trained with the LibriSpeech language modeling resources[2]. The LM is coupled with the acoustic probabilities using beam search and shallow fusion [34]. The LM training is removed from the estimation process. Indeed, it is well-known that neural LMs are extremely energy demanding [6], and it is hence common to use pre-trained models as in our experiments. No LM is used with CommonVoice.

Hyperparameters, and neural architectures vary across the different datasets and are extensively described in the corresponding *SpeechBrain* recipes [32] (commit hash *30e4663*). Pre-trained models are available on *HuggingFace*[3].

4. Energy and CO_2 Estimates

In this section, we first estimate the energy consumption and CO_2 emissions with the experimental setup described in Sec. 3. Then, we analyze the trade-off between performance and carbon cost.

[2]https://www.openslr.org/11/

[3]https://huggingface.co/speechbrain

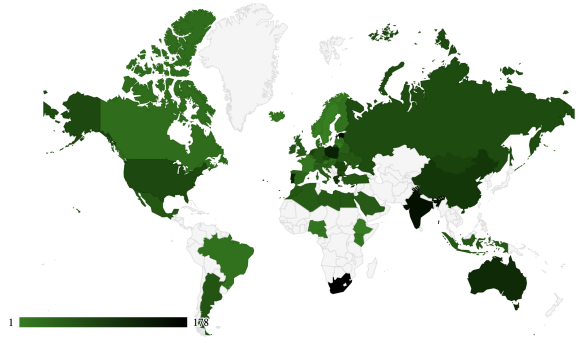

Figure 1: *Global heat map of emitted CO_2 (in kg) to train the best Transformer model on LibriSpeech in different countries. Carbon rates are obtained from governmental sources or [27]. Estimates do not account for varying PUE ($PUE = 1.55$).*

4.1. Training E2E ASR pollutes

The estimates of the energy consumption per iteration and the amount of released CO_2 are reported in Table 2. First, it is important to remark that the obtained WERs are SOTA or competitive with the literature. Indeed, the CRDNN achieves a test WER of 17.70%. To the best of our knowledge, this represents the lowest WER reached so far without self-supervision or pre-training on the CommonVoice-FR dataset. Then, the given LibriSpeech "dev-clean" and "test-clean" results (2.90% and 2.55%) are also comparable with similar systems [35].

As for energy consumption and CO_2 emissions, the main insights that emerge from Table 2 are the following:

- **Geographic location matters.** As shown in numerous studies [6,8], the geographical location has a massive impact on the amount of CO_2 induced by training all ASR models (Figure 1). While France estimates never reach more than 15 kg of CO_2 per model, the same computations in Australia easily emit more than 100 kg.
- **GPU type matters.** It is worth noting the importance of being equipped with recent GPUs. Due to the size of our models, it is often necessary to combine multiple 2080 Ti to match a single Tesla V100. Hence, experiments conducted on CommonVoice with RTX GPUs consumed 2.5 times more energy per epoch compared to Tesla devices.
- **Hyperparameter search is not sustainable**. ASR models are rarely the result of a single training process. For

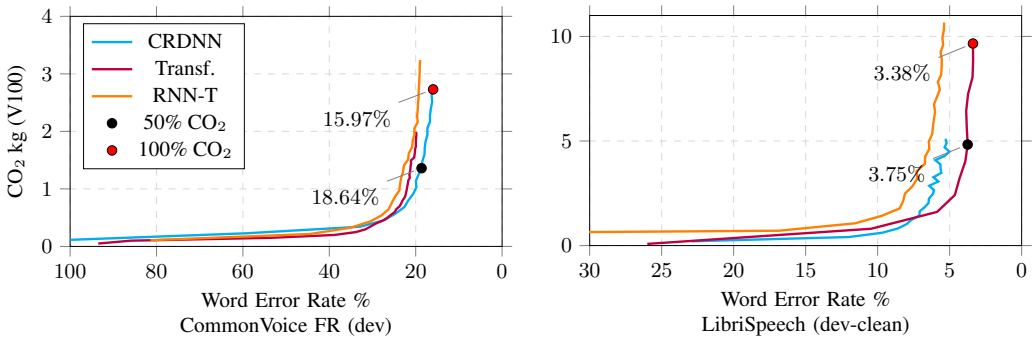

Figure 2: *CO_2 emitted in kg (in France) by different E2E ASR models with respect to the word error rate (WER) on the dev sets of LibriSpeech and CommonVoice. The curves exhibit an exponential trend as most of the training time is devoted to slightly reduce the WER. The black and red dots indicates the WER obtained with* 50% *and* 100% *of the emitted CO_2. On LibriSpeech,* 50% *of the carbon emissions have been dedicated to reach SOTA results with an improvement of* 0.37%.

instance, if we had to run at least 15 experiments before finding a hyperparameter configuration that works reasonably well. This would bring the total amount of CO_2 emitted to 141 kg in the best case and 2055 kg in the worst case (*i.e.*, Transformers trained on LibriSpeech). Such findings are in line with previous observations [10,36,37] advocating for better hyperparameters tuning strategies.

- **All models pollute**. None of the SOTA E2E solutions is really greener than the others. For instance, Transformers process long sequences in parallel due to self-attention resulting in faster epochs. However, Transformers also need larger batch size that slows down the convergence of the system. As a result, Transformer models tend to consume less energy per epoch (0.92 kWh per epoch against 2.00 kWh for RNN-T), but they also require more time to converge to SOTA results hence reducing the gap in terms of CO_2 with the others.

According to the average CO_2 cost of driving a car reported by the European Environment Agency [38], the amount of CO_2 needed to train a single Transformer on LibriSpeech is equivalent to a car trip of 77 km (in the best case) or $1,122$ km (in the worst case) and increase up to $1,155$ km and $16,830$ km with a potential hyperparameters search included (15 runs). Training our greener Transformer also emits as much CO_2 as the entire production and consumption chains of 9.16 kg of bananas [39].

4.2. The unreasonable cost of state-of-the-art performance

Approaching SOTA results is often required to get a novel model accepted by the community. For ASR systems, it becomes crucial to achieve the lowest word error rate possible. Hence, it is important to analyze the trade-off between performance and carbon footprint. Figure 2 depicts the evolution of the WER while training our models versus the corresponding CO_2 emissions.

The curves exhibit an exponential trend, where the amount of CO_2 emissions required to obtain small WER improvements grow very rapidly. Even though this trend was expected, it is important to see how impactful the race for a state-of-the-art model is on the total amount of carbon released. On CommonVoice and with Tesla V100 GPUs, the CRDNN model achieves a WER of 18.64% on the dev set with half of the CO_2 budget. However, doubling this budget only allows the CRDNN to reach a WER of 15.97% on the same set. This represents a gain of 2.67 for a carbon footprint twice larger. This tendency gets even worse on LibriSpeech, a highly competitive benchmark within the speech recognition community. In this context, the Transformer uses 50% of CO_2 emissions to achieve a WER of 3.75% on the dev-clean, and the other 50% of the CO_2 budget to reach the final 3.38%. A gain of 0.37 of WER doubles the carbon bill of the model (4.83 vs 9.66 kg of CO_2). Such a difference certainly does not affect the user perception of the obtained transcription, but it can make the model competitive or not with the SOTA.

These experiments highlight that WER should not be the only metric considered when comparing different ASR models. A more comprehensive evaluation protocol should report WER and energy efficiency.

5. Conclusions

Deep learning might play an increasing role in the climate change tragedy. Fortunately, recent studies started to analyze its environmental cost. In this paper, we extended previous works by estimating CO_2 emissions of SOTA speech recognizers on common benchmarks. We found that the CO_2 emissions generated by these models are far from being negligible. The carbon footprint largely depends on the geographic location, the type of GPU adopted, and increases significantly while performing hyperparameter tuning. Finally, we showed that very minor improvements may be obtained at an extremely high carbon price. Our findings offer a starting point to stimulate the debate around the environmental impact of our research efforts. Moreover, obtained results highlight the need for reducing the practice of accepting or rejecting novel ideas based on the raw performance metric only.

6. Acknowledgements

The authors would like to acknowledge the computing support of Compute Canada and the support from Avignon Université and LIA to compute the PUE of the servers. Thanks to Loren Lugosch and Sasha Luccioni for the valuable comments.

7. References

[1] IPCC, "Climate change 2014 synthesis report," 2014.

[2] R. K. Pachauri, L. Gomez-Echeverri, and K. Riahi, "Synthesis report: summary for policy makers," 2014.

[3] T. J. Crowley, "Causes of climate change over the past 1000 years," *Science*, vol. 289, no. 5477, pp. 270–277, 2000.

[4] D. Rolnick, P. L. Donti, L. H. Kaack, K. Kochanski, A. Lacoste, K. Sankaran, A. S. Ross, N. Milojevic-Dupont, N. Jaques, A. Waldman-Brown, A. Luccioni, T. Maharaj, E. D. Sherwin, S. K. Mukkavilli, K. P. Körding, C. P. Gomes, A. Y. Ng, D. Hassabis, J. C. Platt, F. Creutzig, J. Chayes, and Y. Bengio, "Tackling climate change with machine learning," *CoRR*, vol. abs/1906.05433, 2019.

[5] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green ai," *arXiv preprint arXiv:1907.10597*, 2019.

[6] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in nlp," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3645–3650.

[7] L. F. W. Anthony, B. Kanding, and R. Selvan, "Carbontracker: Tracking and predicting the carbon footprint of training deep learning models," ICML Workshop on Challenges in Deploying and monitoring Machine Learning Systems, July 2020, arXiv:2007.03051.

[8] X. Qiu, T. Parcollet, D. Beutel, T. Topal, A. Mathur, and N. Lane, "Can federated learning save the planet?" in *NeurIPS-Tackling Climate Change with Machine Learning*, 2020.

[9] K. Hao, "Training a single ai model can emit as much carbon as five cars in their lifetimes," *MIT Technology Review*, 2019.

[10] A. Lacoste, A. Luccioni, V. Schmidt, and T. Dandres, "Quantifying the carbon emissions of machine learning," *arXiv preprint arXiv:1910.09700*, 2019.

[11] R. F. Berriel, A. T. Lopes, A. Rodrigues, F. M. Varejão, and T. Oliveira-Santos, "Monthly energy consumption forecast: A deep learning approach," in *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017, pp. 4283–4290.

[12] D. Amodei and D. Hernandez, "Ai and compute," *https://blog.openai.com/aiand-compute*, 2018.

[13] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[14] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *arXiv preprint arXiv:2101.03961*.

[15] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang *et al.*, "A comparative study on transformer vs rnn in speech applications," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 449–456.

[16] J. Li, V. Lavrukhin, B. Ginsburg, R. Leary, O. Kuchaiev, J. M. Cohen, H. Nguyen, and R. T. Gadde, "Jasper: An End-to-End Convolutional Neural Acoustic Model," in *Proc. Interspeech 2019*, 2019, pp. 71–75. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1819

[17] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo, and S. Kumar, "Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7829–7833.

[18] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[19] A. Capozzoli and G. Primiceri, "Cooling systems in data centers: state of art and emerging technologies," *Energy Procedia*, vol. 83, pp. 484–493, 2015.

[20] UptimeInstitue, "2019 data center industry survey results," https://uptimeinstitute.com/2019-data-center-industry-survey-results, 2019.

[21] Google, "Efficiency-data centres," https://www.google.co.uk/about/datacenters/efficiency/, 2020.

[22] AWS, "Aws and sustainability," https://aws.amazon.com/about-aws/sustainability, 2020.

[23] Microsoft, "Datacenter fact sheet - microsoft download center," 2015.

[24] M. Hodak, M. Gorkovenko, and A. Dholakia, "Towards power efficiency in deep learning on data center hardware," in *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019, pp. 1814–1820.

[25] E. E. Agency, "Greenhouse gas emission intensity of electricity generation," European Environment Agency, 2020.

[26] "How much carbon dioxide is produced per kilowatthour of u.s. electricity generation?" U.S Energy Information Administration, 2020.

[27] C. Transparency, https://www.climate-transparency.org global-index.

[28] K. Anderson, "The inconvenient truth of carbon offsets," *Nature*, vol. 484, no. 7392, pp. 7–7, 2012.

[29] P. Bertoldi and T. Huld, "Tradable certificates for renewable electricity and energy savings," *Energy policy*, vol. 34, no. 2, pp. 212–222, 2006.

[30] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[31] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 4218–4222.

[32] M. Ravanelli, T. Parcollet, A. Rouhe, P. Plantinga, E. Rastorgueva, L. Lugosh, N. Dawalatabad, C. Ju-Chieh, A. Heba, F. Grondin, W. Aris, C.-F. Liao, S. Cornell, S.-L. Yeh, H. Na, Y. Gao, S.-W. Fu, C. Subakan, R. De Mori, and Y. Bengio, "Speechbrain," https://github.com/speechbrain/speechbrain, 2021.

[33] A. Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[34] A. Kannan, Y. Wu, P. Nguyen, T. N. Sainath, Z. Chen, and R. Prabhavalkar, "An analysis of incorporating an external language model into a sequence-to-sequence model," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1–5828.

[35] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2613–2617. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2680

[36] P. Henderson, J. Hu, J. Romoff, E. Brunskill, D. Jurafsky, and J. Pineau, "Towards the systematic reporting of the energy and carbon footprints of machine learning," *arXiv preprint arXiv:2002.05651*, 2020.

[37] D. Rolnick, P. L. Donti, L. H. Kaack, K. Kochanski, A. Lacoste, K. Sankaran, A. S. Ross, N. Milojevic-Dupont, N. Jaques, A. Waldman-Brown *et al.*, "Tackling climate change with machine learning," *arXiv preprint arXiv:1906.05433*, 2019.

[38] E. E. Agency, "Monitoring of co2 emissions from passenger cars," 2020.

[39] A. Iriarte, M. G. Almeida, and P. Villalobos, "Carbon footprint of premium quality export bananas: case study in ecuador, the world's largest exporter," *Science of the total environment*, vol. 472, pp. 1082–1088, 2014.