Introduction

Environmental Challenges

• **Computational Demand of Large ML Models:** As ML models, especially in NLP, increase in scale, they become more accurate and capable but also demand more computing power and energy. (NLP models such as T5, Meena, GShard, Switch Transformer, and GPT-3.)

• Energy usage and CO2e as critical metrics for evaluating ML models

**Design New Standard:**

• Improve transparent about energy use and CO2e in their future research

• To include energy usage during training and inference in industry standard benchmarks.
Introduction

Recommendations for the ML Community

• Encourage more researchers to measure energy usage and CO2e—or to get a rough estimate and publish the data
• Efficiency should be an evaluation criterion for publishing ML research on computationally intensive models
• Reducing training time matters: To inspire progress in reducing training costs
Energy Consumption and Carbon Footprint of an NLP Model

The factors that influence an ML model's electricity requirement

- Algorithm
- Implementation Program
- Number of Processors
- Speed and Power of Processors
- Datacenter Efficiency
- Energy Supply Mix: The mix of energy sources used (e.g., renewable, gas, coal, etc.) to power the datacenter also affects the overall electricity usage
Energy Consumption and Carbon Footprint of an NLP Model

\[ \text{Footprint} = (\text{electrical energy}_{\text{train}} + \text{queries} \times \text{electrical energy}_{\text{inference}}) \times \frac{\text{CO2e}_{\text{datacenter}}}{\text{KWh}} \]

- \text{electrical energy}_{\text{train}}: The electrical energy consumed during the training of the ML model.
- \text{queries} \times \text{electrical energy}_{\text{inference}}: The product of the number of queries processed by the model and the electrical energy used per query during inference.
- \frac{\text{CO2e}_{\text{datacenter}}}{\text{KWh}}: The carbon dioxide equivalent emissions per kilowatt-hour specific to the datacenter where the ML operations are performed.
Energy Consumption and Carbon Footprint of an NLP Model

Energy usage in serving vs. training

- **Higher Energy Consumption in Serving**: Most companies spend more energy on serving (performing inference) a DNN model than on training it. (NVIDIA estimated that 80–90% of the ML workload is inference processing. Similarly, Amazon Web Services claimed that 90% of the ML demand in the cloud is for inference.)

- **Energy Distribution between Training and Serving**: If the total energy used in ML is split, approximately 10% goes into training and 90% into serving. This means that even if a given ML model required double the energy for training, it could still reduce overall carbon emissions if the energy used in serving is cut by 20%.

**Easier Assessment of Training Energy Usage**: The energy usage during training is more isolated and thus easier to investigate compared to inference. However, it's important to remember that the carbon footprint of inference is still significant and plays a major role in the overall environmental impact of ML models.
**Datacenter**: The net emissions are significantly lower at the Google location, suggesting more sustainable energy use.

**Datacenter PUE**: The Google location showing a more efficient (lower) PUE of 1.11 compared to the US average of 1.59.

**Training Metrics**: The Google data center is the most efficient, with the lowest training time, energy consumption, and CO2e emissions.

**CO2e Emissions**: Gross and net CO2e emissions for model training are provided, highlighting the environmental impact of the training process. The Google has lower emissions, and the use of TPUs further decreases this footprint.

**Carbon-Free Energy**: The Google data center uses 78% net carbon-free energy.
By optimizing the ML model training process (model selection, processing hardware and data center energy efficiency), carbon emissions can be significantly reduced.

|                          | Small Unit                              | Large Unit                                      |
|--------------------------|-----------------------------------------|------------------------------------------------|
| Energy Consumption       | Kilowatt hours (KWh)                    | Megawatt hours (MWh = 1000 KWh)                |
| Carbon Footprint (CO₂e or CO₂) | Kilograms (kg)                          | Metric ton (t = 1000 kg)                       |
| Perspective (see Appendix A) | Single passenger round trip SF-NY (1.2t CO₂e) | Passenger jet plane round trip SF-NY (180t CO₂e) |
Improvement

Algorithm/Program Improvement

• **Achieving Same Accuracy with Less Computation**: This reduction in computational demand directly translates to lower energy usage.

• **Using Pre-trained Models for Efficiency**: This method leverages existing computational work to reduce further energy expenditure.

• **Techniques for Reducing Energy and Carbon Emissions**:
  • **Distillation**: Reducing the energy required for training and inference.
  • **Pruning, Quantization and Efficient Coding**: Pruning removes unnecessary connections in the network, quantization reduces the precision of the calculations, and efficient coding optimizes the way data is stored and processed.
Improvement

Processor Improvement

- **Increased Speed**: This significant increase in processing speed implies that models can be trained and served in a fraction of the time, leading to reduced energy usage over the duration of these processes.

- **Reduced Power Consumption**: Lower power consumption directly translates to reduced energy requirements for training and inference.

- **Enhanced Performance per Watt**: This improvement indicates a more efficient use of energy, yielding higher computational output for each unit of energy consumed.
Improvement

Datacenter Improvement

• The efficiency of datacenters is measured by the energy overhead metric PUE (Power Usage Effectiveness), which has improved over the years.

• Google's datacenter PUE data shows efficiency improvements, implying better energy utilization and reduced CO2e emissions. (In 2020, The US average was 1.59, Google’s was 1.11, a factor of 1.4X better.)

• Despite the increase in datacenter energy consumption globally, the rate is lower than the growth in computing capacity, suggesting some level of efficiency gains.
Improvement

Improving the energy mix

• 24/7 Carbon Free Energy Framework

• Strategic placement of data centers: Google strategically locates data centers where the energy grid is cleaner or where clean energy can be purchased directly.

• Local Grid and Hourly Accounting: Real-time matching of energy consumption with renewable generation further reduces the net CO2/kWh value of its operations.
Energy Usage and CO2e Emissions of Five Recent Large NLP Models
- T5 is a pre-trained language model
- Meena is a multi-turn open-domain chatbot
- GShard is composed of a set of lightweight annotation APIs
- Switch Transformer simplifies the Mixture of Expert (MoE) routing algorithm
- GPT-3 is an autoregressive language model
Discussion

Estimating the cost of neural architecture search (NAS)

- **Discrepancy in Cost Estimates**: Researchers employed small proxy tasks to search for optimal models to save time and money, and then scaled up these models to full size. Due to this misunderstanding, leading to difficulties in correctly estimating the CO2e retrospectively from the NAS paper.

- **Efficiency of Google Datacenters**: One reason for the lower actual cost is the efficiency of Google datacenters. These datacenters use energy more efficiently, which contributes to lower overall costs for running NAS processes.

**Recommendation for Future Research**: This example underscores the need for researchers to measure energy usage and CO2e for computationally intensive projects and report them when practical, rather than relying on retrospective estimates by others.
1. **Base Transformer vs. Big Transformer**: The Big Transformer typically has more parameters and is expected to perform better at the task.

2. **Diminishing Returns**: The graph suggests that performance gains can diminish as the number of parameters grows very large, where additional complexity provides smaller incremental improvements.

3. **Evolution of Model Performance**: Scaling up the model size within the same architecture can lead to better results.
Discussion

Training Process is Important

• **Multiple Attempts Before Final Training:** It is often necessary to make several attempts to set everything up correctly before arriving at the final training run. This means the final training run alone does not reflect the total cost of resources used in the training process.

• **Challenges in Measuring Total Cost:** Accurately accounting for the total resource cost, including these preliminary attempts, is challenging. This difficulty arises because it's hard to improve or assess what you can't measure.

• **Internal Google Initiative:** Google is developing an internal product to record information about the training process, including data provenance. It will help in understanding the full training lifecycle more comprehensively.

• **Open Source Tools for Tracking:** These tools are vital for recording the training process, including the various attempts leading up to the final run. ("experiment-impact-tracker", "CodeCarbon")

• **Need for Transparent Reporting in Research:** This transparency allows others to recreate and verify results, thereby ensuring accuracy and discouraging misreporting of resource usage in computationally intensive ML research.
Discussion

Measurements are more interesting than extrapolations

The comparison shows that while peak performance is often similar across different models and processors, the measured performance can vary significantly.

TPUv3 consistently shows a higher measured system power across different models compared to the V100 for GPT-3.
Measurements are more interesting than extrapolations

GPT3 indicate more computations can be done for the same amount of energy, reflecting higher energy efficiency.

Measured vs peak performance, measured system power vs peak chip power (TDP), and measured vs peak performance/Watt for V100 GPU and TPU v3
Measurements are more interesting than extrapolations

Importantly, the figure highlights that these calculators do not estimate net CO2e, which could be up to ten times lower than the gross estimates. This emphasizes the significance of direct measurement over indirect calculations for accuracy.

“The ML Emissions and Green Algorithms calculators do not estimate net CO2e, which could be up to 10X lower”
Discussion

Improving Energy Efficiency of Machine Learning Models

• Purpose: To maintain accuracy with less computation.

Specific Techniques Mentioned

• **Distillation**: This process involves transferring knowledge from large, complex models to smaller, more efficient ones.

• **Pruning, Quantization, and Efficient Coding**: These methods can enhance the energy efficiency of deep neural networks (DNNs) by three to seven times.

• **Fine-Tuning and Transfer Learning**: These approaches utilize pre-trained models to avoid starting from scratch for each NLP task, thus saving computational resources.

• **Sparsely Activated Mixture-of-Expert-Style Models**: They can significantly reduce computation and energy costs by more than ten times while providing higher accuracy than equivalent dense Transformer or LSTM-based models.
Discussion

Misconceptions:

• Addresses common fallacies regarding data center energy use:
  • **Full Utilization Fallacy**: The assumption that data centers are always fully utilized is incorrect. Data centers are designed to handle peak demands and may have idle capacity.
  • **Growth Limitation Fallacy**: Contrary to the belief that cloud centers cannot expand quickly, the text argues that like universities, cloud companies invest in more infrastructure as needed.
  • **Renewable Energy Limitation Fallacy**: It refutes the idea that renewable energy cannot grow. The availability of renewable energy can fluctuate, but companies like Google are investing in the creation of new renewable resources, thus expanding the renewable energy supply.
  • **Resource Competition Fallacy**: The training of NLP models on Google’s ML processors does not compete with other tasks within the data center. Different tasks, like those for CPUs and TPUs, do not interfere with each other.
  • **Latency Requirement Fallacy**: Contrary to the belief that training must run in all data centers, user-facing inference applications need global distribution, but training can be limited to fewer, selected data centers.
  • **No Business Reason Fallacy**: The text challenges the idea that there's no economic incentive to reduce carbon emissions. Google and other companies have been proactive in matching their data center energy usage with renewable energy sources, aiming to achieve a carbon-neutral status.
Discussion

The Availability of Green Datacenters:

- **Question**: Whether such green data centers are available to all ML practitioners, not just those within large organizations like Google.

- **Answer**: Many have access to energy-optimized datacenters reducing the cost of training matters too. The training of those four large NLP models is not a significant fraction of Google’s energy consumption.
Discussion

How does training a large NLP model compare to other activities?

• **Comparison with Air Travel**: The emissions of a direct round trip by a whole passenger jet between San Francisco and New York are estimated at 180 metric tons of CO2 equivalent (tCO2e). Compared to this figure:
  - T5: Approximately 26% of this jet trip.
  - Meena: About 53%.
  - Gshard-600B: Roughly 2%.
  - Switch Transformer: Around 32%.
  - GPT-3: Approximately 305%.

• **Comparison with Bitcoin**: The annual carbon footprint of Bitcoin mining is comparable to 200,000 to 300,000 round trips between San Francisco and New York by passenger jet.

• **Google’s Total Energy Consumption**: Training of these large NLP models is not a significant portion of Google's overall energy usage.
Discussion

Are the benefits of NLP models worth the energy cost?

- The performance of all models improves as the number of examples per language increases, which suggests that having more data improves the quality of machine translation.
- A clear trend is that models with more available parameters perform better in terms of BLEU scores.
Thank you

University of Massachusetts
Amherst