Green Information Retrieval Research

Harry Scells
Leipzig University, Germany
PART I

Context
Why?

• Large (pre-trained) neural language models

• Expend high energy for training and inference (compared to traditional models)

• The energy demands expected to continue growing as size and complexity of models increase

• Data centers and other infrastructure used to run these models also consume energy
What about IR research?

But what are emissions?

- **Energy**: amount of work done
  - Measured in **joules**
But what are emissions?

- **Energy**: amount of work done
  - Measured in **joules**

- **Power**: energy per unit time
  - Measured in **watts**; 1 watt = 1 joule/second
  - kWh: energy consumed at a rate of 1 kilowatt for 1 hour
But what are emissions?

- **Energy**: amount of work done
  - Measured in **joules**

- **Power**: energy per unit time
  - Measured in **watts**; 1 watt = 1 joule/second
  - kWh: energy consumed at a rate of 1 kilowatt for 1 hour

- **Emissions**: by-products created by producing power
  - Measured in **kgCO₂e**; kilograms of carbon dioxide equivalent
What about IR research?

Isn’t this just retrieval efficiency?

Retrieval Efficiency

- **Speed** a system is able to retrieve relevant documents or information in response to a query.

- Factors that can impact retrieval efficiency include:
  - Size and complexity of the corpus being searched
  - Effectiveness of the retrieval models or techniques being used
  - Efficiency of the hardware and infrastructure used
Effectiveness vs Efficiency

- Happy
- Sad
Effectiveness vs. Efficiency:

- High effectiveness with high efficiency.
- High effectiveness but lower efficiency.
- Low effectiveness with higher efficiency.
Effectiveness

Efficiency

Utilisation
Okay, so what does this mean for IR?
Utilisation and Green IR

**Green IR is...**

- “research that yields novel results while taking into account the computational cost, encouraging a reduction in resources spent” [2]

Utilisation and Green IR

Green IR is...
- “research that yields novel results while taking into account the computational cost, encouraging a reduction in resources spent” [2]
- Neural methods require pre-trained LMs
- **Expensive** to create
- Trend in IR towards creating **IR-specific** LMs [3,4,5,6]

Utilisation and Green IR

Green IR is...
- “research that yields novel results while taking into account the computational cost, encouraging a reduction in resources spent” [2]
- Neural methods require pre-trained LMs
- **Expensive** to create
- Trend in IR towards creating **IR-specific** LMs [3,4,5,6]

Pre-trained LMs come at a high power and emissions cost

---

Utilisation and Green IR

Green IR is...

- "research that yields novel results while taking into account the computational cost, encouraging a reduction in resources spent" [2]

- Neural methods require pre-trained LMs
- **Expensive** to create
- Trend in IR towards creating **IR-specific** LMs [3,4,5,6]

Pre-trained LMs come at a high power and emissions cost

- Missing dimension of IR evaluation
  - Effectiveness
  - Efficiency
  - **Utilisation**

Okay, so how can I measure this?

Okay, so what does this mean for IR?

Okay, so how can I measure this?
Measuring emissions

• First, measure power consumption:
Measuring emissions

• First, measure power consumption:

\[ p_t = \frac{\Omega \cdot t \cdot (p_c + p_r + p_g)}{1000} \]
Measuring emissions

• First, measure power consumption:

\[ p_t = \frac{\Omega \cdot t \cdot (p_c + p_r + p_g)}{1000} \]
Measuring emissions

• First, measure power consumption:

\[ p_t = \Omega \cdot t \cdot \frac{(p_c + p_r + p_g)}{1000} \]
Measuring emissions

• First, measure power consumption:

\[ \Omega \cdot t \cdot (p_c + p_r + p_g) \frac{\text{watts}}{1000} \]
Measuring emissions

• First, measure power consumption:

\[ p_t = \frac{\Omega \cdot t \cdot (p_c + p_r + p_g)}{1000} \]

- PUE
- Running Time
- CPU, RAM, GPU power draw
- watts
Measuring emissions

• First, measure power consumption:

\[ P_t = \frac{\Omega \cdot t \cdot (p_c + p_r + p_g)}{1000} \]

watts

PUE

Running Time

CPU, RAM, GPU power draw

• Next, measure emissions:
Measuring emissions

• First, measure power consumption:

\[ p_t = \frac{\Omega \cdot t \cdot (p_c + p_r + p_g)}{1000} \]

• Next, measure emissions:

\[ \text{kgCO}_2\text{e} = \theta \cdot p_t \]
Measuring emissions

• First, measure power consumption:

\[ p_t = \frac{\Omega \cdot t \cdot (p_c + p_r + p_g)}{1000} \]

• Next, measure emissions:

\[ \text{emissions} \rightarrow \text{kgCO}_2e = \theta \cdot p_t \]
Measuring emissions

• First, measure power consumption:

\[ p_t = \frac{\Omega \cdot t \cdot (p_c + p_r + p_g)}{1000} \]

• Next, measure emissions:

\[ \text{kgCO}_2\text{e} = \theta \cdot p_t \]
Measuring emissions

• First, measure power consumption:

$$ p_t = \frac{\Omega \cdot t \cdot (p_c + p_r + p_g)}{1000} $$

• Next, measure emissions:

$$ \text{kgCO}_2\text{e} = \theta \cdot p_t $$

avg. CO$_2$e (kg) per kWh where experiments took place
Measuring emissions

• First, measure power consumption:

\[ p_t = \frac{\Omega \cdot t \cdot (p_c + p_r + p_g)}{1000} \]

• Next, measure emissions:

\[ \text{avg. CO}_2\text{e (kg) per kWh where experiments took place} \]

\[ \text{Power consumption of experiments} \]

• Emissions of my search engine:

\[ \text{kgCO}_2\text{e} = \theta \cdot p_t \]

\[ \text{kgCO}_2\text{e} = \theta \cdot \Delta_q \cdot p_q \]
Measuring emissions

• First, measure power consumption:

\[ p_t = \frac{\Omega \cdot t \cdot (p_c + p_r + p_g)}{1000} \]

• Next, measure emissions:

\[ \text{avg. CO}_2\text{e (kg) per kWh where experiments took place} \]

\[ \text{emissions} \rightarrow \text{kgCO}_2\text{e} = \theta \cdot p_t \]

• Emissions of my search engine:

\[ \text{kgCO}_2\text{e} = \theta \cdot \Delta_q \cdot p_q \]
Measuring emissions

• First, measure power consumption:

\[ p_t = \frac{\Omega \cdot t \cdot (p_c + p_r + p_g)}{1000} \]

• Next, measure emissions:

\[ \text{avg. CO}_2\text{e (kg) per kWh where experiments took place} \]

\[ \text{Power consumption of experiments} \]

• Emissions of my search engine:

\[ \text{No. queries issued per unit time} \]

\[ \text{Power consumption of a single query} \]
# Measuring power and emissions in practice

<table>
<thead>
<tr>
<th>Name</th>
<th>CPU</th>
<th>DRAM</th>
<th>GPU</th>
<th>Network</th>
<th>Repository</th>
</tr>
</thead>
<tbody>
<tr>
<td>pyJoules</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>X</td>
<td><a href="https://github.com/powerapi-ng/pyJoules">https://github.com/powerapi-ng/pyJoules</a></td>
</tr>
<tr>
<td>Cumulator [81]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td><a href="https://github.com/epfl-iglobalhealth/cumulator">https://github.com/epfl-iglobalhealth/cumulator</a></td>
</tr>
</tbody>
</table>

```python
from codecarbon import EmissionsTracker

tracker = EmissionsTracker()
tracker.start()
# Experiment code goes here
tracker.stop()
```
Okay, so what does this mean for IR?
Okay, so how can I measure this?
Okay, so show me what it means in IR research practice!
Experimental Setup Overview

- Methods:
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - uniCOIL
  - TILDEv2
Experimental Setup Overview

- Methods:
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - uniCOIL
  - TILDEv2

Non-neural
Experimental Setup Overview

- Methods:
  - BM25 (Non-neural)
  - LambdaMART (Non-neural)
  - DPR
  - monoBERT ("Neural")
  - uniCOIL ("Neural")
  - TILDEv2 ("Neural")
Experimental Setup Overview

• Methods:
  • BM25
  • LambdaMART
  • DPR
  • monoBERT
  • uniCOIL
  • TILDEv2

Non-neural
Dense retriever (bi-encoder)

Cosine similarity

BERT
Query

BERT
Document
Experimental Setup Overview

- Methods:
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - uniCOIL
  - TILDEv2
Experimental Setup Overview

- **Methods:**
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - uniCOIL
  - TILDEv2

- **Non-neural:**
  - Sparse retrievers

- **Dense retriever (bi-encoder):**
  - BERT (cross-encoder)

- **Scorer**
  - Tokeniser
  - BERT

- **Query**
- **Document**
Experimental Setup Overview

- Methods:
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - uniCOIL
  - TILDEv2

- Dense retriever (bi-encoder)
  - BERT (cross-encoder)

- Sparse retrievers
  - Non-neural

- Process documents offline
Experimental Setup Overview

- Methods:
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - uniCOIL
  - TILDEv2

- Non-neural

- Dense retriever (bi-encoder)
  - BERT (cross-encoder)

- Sparse retrievers

- Query Document
  - Tokeniser
    - Fast inference time (Can even be done on CPU)
  - BERT
    - Process documents offline
Experimental Setup Overview

• Methods:
  • BM25
  • LambdaMART
  • DPR
  • monoBERT
  • uniCOIL
  • TILDEv2

Dense retriever (bi-encoder)
BERT (cross-encoder)
Sparse retrievers

Exact match
Scorer

Tokeniser
Query

Fast inference time
(Can even be done on CPU)

Document

Process documents offline
Document expansion
TILDE/doc2query
Experimental Setup Overview

- **Methods:**
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - uniCOIL
  - TILDEv2

  - **Non-neural**

- **Collection:**
  - MSMARCOv1

  - **Experiments:**
Experimental Setup Overview

- **Methods:**
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - uniCOIL
  - TILDEv2

- **Collection:**
  - MSMARCOv1

- **Experiments:**
  - How many emissions do these methods produce to obtain an experimental result?
Experimental Setup Overview

- **Methods:**
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - uniCOIL
  - TILDEv2

- **Collection:**
  - MSMARCOv1

- **Experiments:**
  - How many emissions do these methods produce to obtain an experimental result?
  - What are the effectiveness-utilisation trade-offs of these methods?
Experimental Setup Overview

- **Methods:**
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - uniCOIL
  - TILDEv2

- **Collection:**
  - MSMARCOv1

- **Experiments:**
  - How many emissions do these methods produce to obtain an experimental result?
  - What are the effectiveness-utilisation trade-offs of these methods?
How many emissions do these methods produce to obtain an experimental result?

- BM25
- LambdaMART
- DPR
- monoBERT
- TILDEv2
- uniCOIL

<table>
<thead>
<tr>
<th>Method</th>
<th>Emissions (kgCO₂)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM25</td>
<td>0.00168</td>
</tr>
<tr>
<td>LambdaMART</td>
<td>0.00190</td>
</tr>
</tbody>
</table>
How many emissions do these methods produce to obtain an experimental result?

- Methods:
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - TILDEv2
  - uniCOIL
How many emissions do these methods produce to obtain an experimental result?

- Methods:
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - TILDEv2
  - uniCOIL
How many emissions do these methods produce to obtain an experimental result?

- Methods:
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - TILDEv2
  - uniCOIL
How many emissions do these methods produce to obtain an experimental result?

- BM25
- LambdaMART
- DPR
- monoBERT
- TILDEv2
- uniCOIL
- uniCOIL+TILDE
- TILDEv2+TILDE
- uniCOIL+doc2query
- TILDEv2+doc2query
How many emissions do these methods produce to obtain an experimental result?

Neural methods produce considerably more emissions than non-neural.

![Graph showing emissions comparison]

- Methods:
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - TILDEv2
  - uniCOIL
How many emissions do these methods produce to obtain an experimental result?

- Document expansion can have big impact on emissions

Neural methods produce considerably more emissions than non-neural

- Methods:
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - TILDEv2
  - uniCOIL

Emissions (kg\(\text{CO}_2\))

- Neural methods produce considerably more emissions than non-neural
How many emissions do these methods produce to obtain an experimental result?
How many emissions do these methods produce to obtain an experimental result?
How many emissions do these methods produce to obtain an experimental result?
What are the effectiveness-utilisation trade-offs of these methods?
What are the effectiveness-utilisation trade-offs of these methods?

<table>
<thead>
<tr>
<th>Method</th>
<th>Emissions (kgCO2e)</th>
<th>Effectiveness (MRR@10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>uniCOIL+doc2query</td>
<td>0</td>
<td>0.185</td>
</tr>
<tr>
<td>uniCOIL+TILDE</td>
<td>20</td>
<td>0.209</td>
</tr>
<tr>
<td>TILDEv2+TILDE</td>
<td>40</td>
<td>0.234</td>
</tr>
<tr>
<td>monoBERT</td>
<td>60</td>
<td>0.258</td>
</tr>
<tr>
<td>uniCOIL+TILDE</td>
<td>80</td>
<td>0.283</td>
</tr>
<tr>
<td>TILDEv2+doc2query</td>
<td>100</td>
<td>0.307</td>
</tr>
<tr>
<td>monoBERT</td>
<td>120</td>
<td>0.331</td>
</tr>
<tr>
<td>DPR</td>
<td>140</td>
<td>0.356</td>
</tr>
<tr>
<td>BM25</td>
<td>0</td>
<td>0.38</td>
</tr>
</tbody>
</table>
What are the effectiveness-utilisation trade-offs of these methods?

The chart illustrates the trade-offs between effectiveness (MRR@10) and emissions (kgCO2e) for various methods:

- uniCOIL+TILDE
- uniCOIL+doc2query
- TILDEv2+TILDE
- TILDEv2+doc2query
- monoBERT
- BM25
- DPR
- BM25

The arrow indicates an increase in utilisation with a corresponding decrease in effectiveness.
What are the effectiveness-utilisation trade-offs of these methods?

Effectiveness (MRR@10)

- uniCOIL+TILDE
- TILDEv2+TILDE
- monoBERT
- uniCOIL+doc2query
- TILDEv2+doc2query
- BM25

Emissions (kgCO2e)

- More utilisation = Higher effectiveness
What are the effectiveness-utilisation trade-offs of these methods?

Effectiveness (MRR@10) vs Emissions (kgCO2e)

More utilisation = Higher effectiveness
What are the effectiveness-utilisation trade-offs of these methods?
PART II
Green IR in Practice
A framework for practitioners to remain mindful of potential costs of IR research
Reduce

Vs

JAM

JAM
Reduce

Vs

Expend fewer resources
Reduce

- Straightforward: simply reduce the number of experiments

- Limit expensive computations, e.g., use CPU, FPGAs over GPU

- Prior to starting any research or experiments, ask: *How can I perform research with fewer resources?*
  
  - Random hyper-parameter search
  
  - CPU-based inference
Reuse
Reuse

Repurpose resources intended for one task to the same task
Reuse

• Reuse existing software artefacts such as data, code, or models

• Reuse: take something existing and repurpose it for the same task it was devised for

• Prior to starting any research or experiments, ask: *How can I repurpose data, code, or other digital artefacts meant for one task to the same task?*

  • Reuse large collections

  • Pre-indexing common collections
Recycle

[Image: Jar of jam on the left, candle on the right]
Recycle

Repurpose resources intended for one task to a different task
Recycle

- Recycle existing software artefacts such as data, code, or models

- Recycle: the action of repurposing an existing artefact for a task it was not originally intended for

- Prior to starting any research or experiments, ask: *How can I repurpose existing data, code, or other digital artefacts meant for one task to a different task?*

  - Neural query expansion

  - Passage expansion with models like TILDE
reduce, reuse, recycle

- Reduce: Expend fewer resources
- Reuse: Repurpose resources intended for one task to the same task
- Recycle: Repurpose resources intended for one task to a different task
PART III

Summary
Efficiency is not just query latency

- There is a trend of “query efficient” neural models which move the heavy computation offline.
- This computation still costs: time, hardware, energy, emissions.
- It is not just a “once off” cost.
Efficiency is not just latency, energy

- Data efficiency
- Learning with little data
- Frugal models, federated learning, few-shot, zero-shot, prompt learning
Summary

• **Larger neural models** = power-hungry hardware = utilisation of more power
  • However: increased model size for higher effectiveness may not apply to IR, as it does to NLP and ML
Summary

• Larger neural models = power-hungry hardware = utilisation of more power
  • However: increased model size for higher effectiveness may not apply to IR, as it does to NLP and ML

• Likely trend in neural IR: go beyond PLMs designed for NLP but are specialised for IR… pre-train for IR
  • More power and more emissions
  • DSI: end-to-end transformers that encapsulate the entire indexing and searching architecture into a single model
Summary

- **Larger neural models** = power-hungry hardware = utilisation of more power
  - However: increased model size for higher effectiveness may not apply to IR, as it does to NLP and ML

- Likely trend in neural IR: go beyond PLMs designed for NLP but are specialised for IR… **pre-train for IR**
  - More power and more emissions
  - DSI: end-to-end transformers that encapsulate the entire indexing and searching architecture into a single model

- IR community at a **turning point**
  - Bigger/more complex models
  - Bigger collections of documents, queries
Summary

• **Larger neural models** = power-hungry hardware = utilisation of more power
  • However: increased model size for higher effectiveness may not apply to IR, as it does to NLP and ML

• Likely trend in neural IR: go beyond PLMs designed for NLP but are specialised for IR… **pre-train for IR**
  • More power and more emissions
  • DSI: end-to-end transformers that encapsulate the entire indexing and searching architecture into a single model

• IR community at a **turning point**
  • Bigger/more complex models
  • Bigger collections of documents, queries

• There is a cost to IR (+NLP, ML) research:
  • Power usage: $$$
  • Emissions: CO$_2$e