Paper Presentation on:

GREEN AI

Roy Schwartz, Jesse Dodge, Noah A. Smith, Oren Etzion

CMSC 491/691-11 - Interactive Fiction and Text Generation, Fall 24

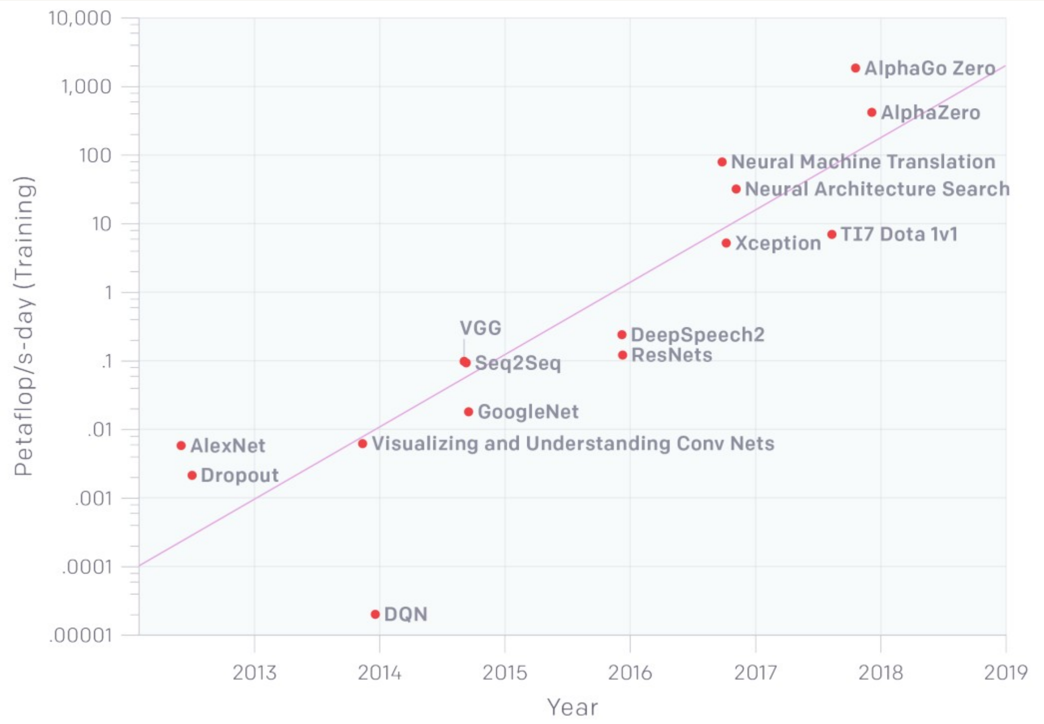
By:

Sukhbir Singh Sardar
MM81414

UMBC

INTRODUCTION

- Current focus on accuracy



- Why Only Accuracy?

- Impact on Accessibility

  High computational power and resources, creates a barrier for students and researchers with limited access to high-end hardware.

- An example where accuracy is given importance

  (ref: *https://leaderboard.allenai.org/robothor_objectnav/submissions/public*)

Rank	Submission	Created	Test SPL	Val SPL	Test Success Rate	Val Success Rate	Test SPL (Proximity)	Val SPL (Proximity)	Test Success Rate (Proximity)	Val Success Rate (Proximity)
1	ProcTHOR Fine-Tune *Anon*	05/02/2022	**0.2884**	**0.2744**	**0.6515**	**0.6639**	**0.2882**	**0.2748**	**0.6534**	**0.6672**
2	ProcTHOR EmbCLIP Zero-Shot (R... *Anonymous submission*	09/27/2023	0.2424	0.2390	0.5132	0.4944	0.2413	0.2411	0.5118	0.4994
3	ProcTHOR Zero-Shot *Anon*	04/27/2022	0.2374	0.2220	0.5495	0.5133	0.2366	0.2216	0.5500	0.5156
4	ProcTHOR EmbCLIP-Codebook Zer... *Anonymous submission*	09/27/2023	0.2365	0.2146	0.5500	0.5111	0.2338	0.2141	0.5490	0.5139
5	EmbCLIP (AI2 @ CVPR'22) *Apoorv Khandelwal, Luca Weihs...*	11/06/2021	0.2004	0.2599	0.4701	0.5222	0.2035	0.2644	0.4760	0.5300
6	ICT-ISIA-Action-boost-model *Yubing Bai, Sixian Zhang, Wei...*	05/23/2021	0.1156	0.1749	0.2750	0.3661	0.1225	0.1783	0.2951	0.3761
7	RGB+D ResNet18-ImageNet *AI2 Team*	03/02/2021	0.1133	0.1737	0.2632	0.3511	0.1205	0.1818	0.2833	0.3700
8	random_submit *random*	03/27/2021	0.1126	0.1750	0.2667	0.3583	0.1204	0.1803	0.2882	0.3761
9	ICT-ISIA-Baseline-with-detect... *Yubing Bai, Sixian Zhang, Wei...*	05/27/2021	0.1118	0.1765	0.2662	0.3828	0.1162	0.1826	0.2804	0.3972
10	50% Random + 50% Pretrained *Baseline*	03/02/2021	0.0618	0.0956	0.1373	0.1867	0.0748	0.1072	0.1647	0.2072
11	Random *Baseline*	03/02/2021	0.0161	0.0133	0.0201	0.0172	0.0433	0.0363	0.0647	0.0561

RED AI

- Red AI?
  AI that prioritizes "state-of-the-art accuracy" by using massive computational resources.

- Achieving better accuracy comes with exponentially higher costs in computation.

- Linear performance gains require exponentially larger models and more experiments.

- For estimating Computational costs for a model, authors made this relation:

$$Cost(R) \propto E \cdot D \cdot H$$

The cost of an AI (R)esult grows linearly with the cost of processing:

- a single (E)xample,
- the size of the training (D)ataset and
- the number of (H)yperparameter experiments.

- Hence, this increases the cost with increasing number of Examples, Datasets and Hyperparameters experiments. (some more parameters like epochs will also increase cost)

- Authors suggested 2 ways AI researchers can focus on efficiency as well as accuracy:

1. Accuracy vs. Training Set Size: promotes data-efficient research.
2. Accuracy vs. Computational Cost: helps evaluate if the accuracy improvements justify the energy and resource costs.

GREEN AI

- Green AI?

  Research that give novel results without increasing computational cost, ideally reducing it.

  Recognizing efficiency alongside accuracy allows researchers to focus on GREENER, more inclusive AI models, benefiting both the environment and accessibility.

- But how do you measure efficiency of model?

  Efficiency measures should allow for fair comparisons between models, regardless of lab, time, or hardware differences. Some of the measures are listed below:

1. CARBON EMISSION ✗

- Impractical to measure the exact amount of carbon
- Not comparable between researchers in different locations or even the same location at different times.

2. ELECTRICITY USAGE ✗

- GPUs often report the amount of electricity
- This measure is hardware dependent, and as a result does not allow for a fair comparison

METHODS OF EFFICIENCY

3. ELAPSED REAL TIME

- A faster model does less computational work.
- Influenced by Hardware, other jobs running on the same machine, and the number of cores used.
- Hence, hinder the comparison between different models.

4. NUMBER OF PARAMETERS

- Does not depend on the underlying hardware.
- But different algorithms make different use of their parameters.
- Hence different models perform different amount of work.

WHAT SHOULD YOU CONSIDER THEN?

METHODS OF EFFICIENCY

5. FLOATING POINT OPERATIONS (FPO

- Provides an estimate to the amount of work performed by a computational process.
- It is computed analytically by defining a cost to two base operations, ADD and MUL.
- It is Recursive function of these two operations for any machine learning abstract operation.

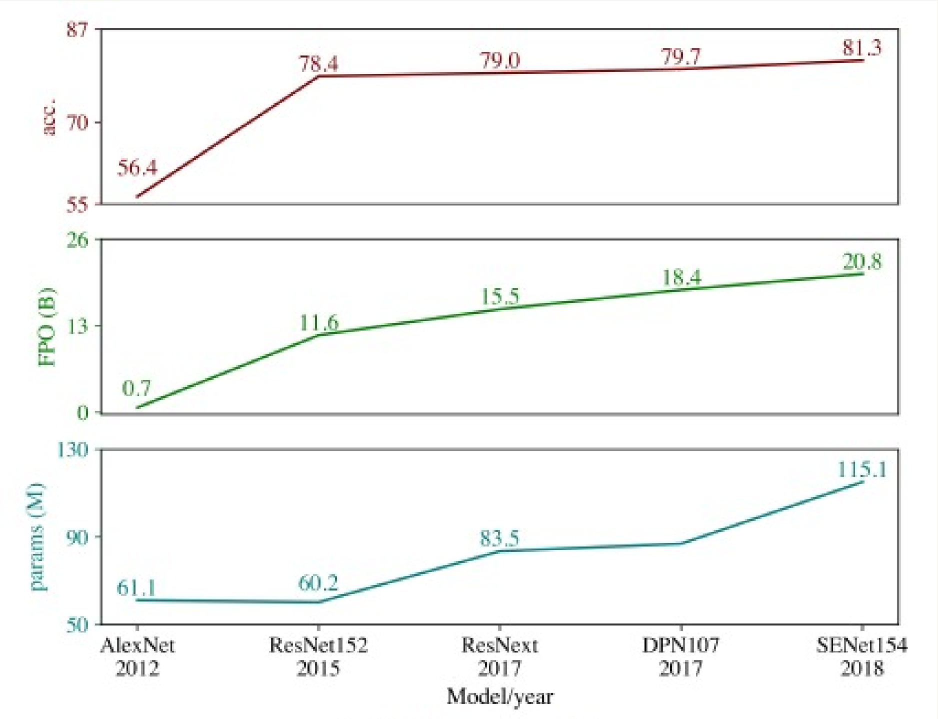
But how do we know it works well?

1. It directly calculates the amount of work done by the machine, closely linking it to the energy consumed during model execution.

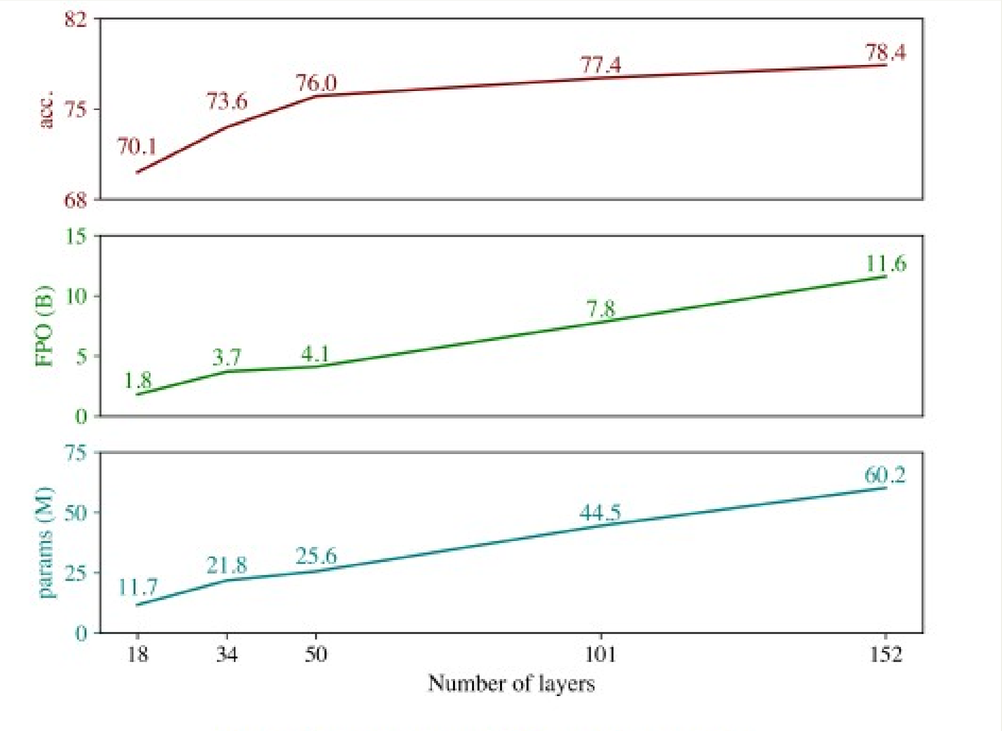
2. FPO is independent of the hardware used, enabling fair comparisons across different models and approaches.

3. FPO is closely linked to the model's running time and accounts for work done at each step, unlike asymptotic runtime.

FPO COSTS ON EXISTING MODELS



(a) Different models.



(b) Different layers of the ResNet model.

LIMITATIONS OF FPO:

- FPO focuses on measuring electricity usage but overlooks other factors like memory consumption, which can increase energy and costs.
- The amount of work a model performs can vary greatly based on its implementation, even for the same model, leading to different processing requirements.

CORPORATE SHIFTS TOWARDS EFFICIENCY:

Companies like Google and Microsoft are prioritizing sustainability, even planting trees to offset their energy consumption.

STRENGTHS AND CHALLENGES OF GREEN AI

STRENGTHS:

- Conveys environmental impacts of heavy AI models.
- Promotes wider participations in AI research by decreasing resource shortages, that enables more researchers (especially students) to contribute to the field of AI.
- Encourages using efficiency measures like FPO to evaluate model performance, aiming for a more balanced and fair way to assess AI beyond just accuracy.

CHALLENGES:

- Complex to also consider efficiency along with accuracy.
- The FPO metric is not being yet used by researches, which would be a slow progress.

GREEN AI IN STORY GENERATION AND INTERACTIVE FICTION

- A real-time story generation model would respond more faster to user inputs when Green AI is taken into consideration.
- As you focus more on making a making a model efficient and smaller, it makes interactive fiction creation more accessible to the users having limited resources (especially students).
- The model can be made such that they allow us to create complex and interactive narratives without requiring strong computational resources.
- Following the policies of Green AI, makes it easier to create personalized game content and stories for each player, all while using fewer resources and being more efficient.

CONCLUSION

WHAT WE HAVE COVERED:

- Scenarios of AI researchers focusing more on accuracy
- AI vs AI
- Methods to measure Efficiency
- A special metric to measure cost, FPO
- Strengths and Weaknesses of Green AI
- Applications of Green AI in Interactive Fiction and Automated Story Generation

Thank you!

ANY QUESTIONS?

Sukhbir Singh Sardar
MM81414
ssardar1@umbc.edu