Emerging AI Technology Trends
Shubhashis Sengupta
accenture
Introduction

5 Key Technology Innovation Trends are shaping the future of AI engineering and Adoption

- Large Language and Image Learning Models
- Gaining Efficiency and Sustainability in AI
- Synthetic Data comes to the Rescue
- AI and Intelligence
- AI Automation at Scale
TREND1: Large-Scale “Meta-Learning” Models

- Transformer models – learning from Terabytes of data and Billions of Parameters
- 3 key Innovations –
  - Attention
  - Self Attention
  - Positional Embedding
- MIT calls these Foundation Models

- Usage and adaptation are changing dramatically –
  - Few Shot and Zero-Shot learning through “Smart Prompts”
  - Fine Tuning of models through LLRD, SWA, Adaptive Tokenization, LoRA
  - Model access, open sourcing and BERTOLOGY
- Meta Learning as models learn representations for multiple tasks (summarization, translation, Q&A, classification, lang generation, sentiment analysis etc.)

- Similar advancement in Image Language Modeling and Generation –
  - DALL-E, DALL-E2
  - Multi-modal Transformers

Image courtesy: Cross Validated, Stack Exchange

Image courtesy: https://twitter.com/OpenAI/

Image courtesy: lilianweng.github.io

Image courtesy: ISO / IEC AI Workshop | 24 – 25 May 2022

SC 42 – Artificial Intelligence

accenture
TREND2: Efficient AI (Faster, Cheaper, More Sustainable)

- Large Models are Costly and not Environment friendly
  - Google Switch (1.6 T params) takes 1 MUSD to train and emits 284T of Co2e
  - Training GPT-3 once may take up to 800K USD

- Making AI training and Optimization efficient at Data Centers
  - Algorithmic Optimization – Pruning and Clustering
  - Deployment Optimization – Equalization, Fold-Batch norms, Fused layers, Quantization

- Training speedups: FFT models, Pathway Dataflow for parallelism

- How to induce Pruning / Sparsity without Model Degradation

<table>
<thead>
<tr>
<th>Model</th>
<th>Hardware</th>
<th>Power (W)</th>
<th>Hours</th>
<th>kWh/PUE</th>
<th>CO2e</th>
<th>Cloud compute cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transformed</td>
<td>P100x8</td>
<td>1415.78</td>
<td>12</td>
<td>27</td>
<td>26</td>
<td>$415-$64</td>
</tr>
<tr>
<td>Transformed</td>
<td>P100x16</td>
<td>1515.43</td>
<td>84</td>
<td>201</td>
<td>192</td>
<td>$289-$3914</td>
</tr>
<tr>
<td>ELMo</td>
<td>P100x3</td>
<td>517.66</td>
<td>336</td>
<td>275</td>
<td>262</td>
<td>$435-$1472</td>
</tr>
<tr>
<td>BERT-base</td>
<td>V100x64</td>
<td>12,041.51</td>
<td>79</td>
<td>1507</td>
<td>1458</td>
<td>$3751-$12,571</td>
</tr>
<tr>
<td>BERT-base</td>
<td>TPU/v2x16</td>
<td>96</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>$3074-$6,012</td>
</tr>
<tr>
<td>NAS</td>
<td>P100x8</td>
<td>1515.43</td>
<td>274,120</td>
<td>656,347</td>
<td>626,155</td>
<td>$942,973-$3,201,722</td>
</tr>
<tr>
<td>NAS</td>
<td>TPU/v2x1</td>
<td>—</td>
<td>32,023</td>
<td>—</td>
<td>—</td>
<td>$44,055-$146,848</td>
</tr>
<tr>
<td>GPT-2</td>
<td>TPU/v2x32</td>
<td>168</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>$12,000-$43,008</td>
</tr>
</tbody>
</table>

Table 3: Estimated cost of training a model in terms of CO2 emissions (lbs) and cloud compute cost (US$). Power and carbon footprint are credited for TPUs due to lack of public information on power draw for this hardware.

- Efficient Edge Computing Architecture
  - TinyML models for low AI footprint
  - Optimized models like MobileNet V1 (DS-CNN), V2
  - Neuromorphic Architecture (SNN) – extremely low power architecture

Source: https://blog.learningtree.com/carbon-footprint-ai-deep-learning/

https://www.tinyml.org/

Spiking Neural Network: https://en.wikipedia.org/wiki/Spiking_neural_network

Image Courtesy: Qualcomm
TREND3: Synthetic Data for AI

- **Right Data is the most Difficult Ask**
  - Getting right domain data for Model Training is getting increasingly difficult
    - GDPR Regulation, Sensitive (PHI, PII) data, Localization and Fiduciary implications, Data Security

- **Open Source and Linked Data exploitation has hit limits**

- **Data Generation for AI**
  - Data De-identification
  - Data Augmentation
  - Fully Synthetic Generative Modeling
    - Models with Implicit likelihood (GAN -> generates by comparison)
    - Models with Explicit likelihood (VAE, Fully Observable Models)
    - Diffusion Network (very realistic)

- **Quality and Trustworthiness**
  - Should not copy but preserve relationship/correlation
  - Statistically meaningful
  - Predictivity, Diversity, Realism, Privacy Preservation

Source: ICML Tutorial on Synthetic Data Generation (https://www.youtube.com/watch?v=_EEH9HU2EE0)
TREND 4: Cognitive AI (Can AI augment / partner with Humans?)

- Narrow AI supremacy (already achieved)
  - AI is already performing at "superhuman" levels for many specific jobs (Q&A, Reading Comprehension, Image Segmentation)
  - Generalization to multi-task level is the key challenge

- Towards Complex and Creative Tasks (In Progress)
  - Planning and Strategy (Games) – Deep Reinforcement Learning (AlphaZero) vs Rules
  - Common Sense Reasoning – Atomic
  - Analogy based Reasoning, Forward Chaining
  - NLI / NLE
  - Multi-step Reasoning with Explanation (logically solving Math Words Problem) – Google PaLM
  - Code generation - CoPilot
  - KG based reasoning (link prediction, message passing)
  - Image Incongruity Detection
  - Winograd Schema

Achieving AI Singularity? Not Yet

TREND5: AI Automation at scale; AI as a Service (AAAS)

- **Cloud-based pre-trained and (slightly) trainable models**
  - Easy to set up, operate and consume
  - SaaS - Mainly driven by AutoML, templates, large pre-trained models.
  - PaaS / Workbench facilities – SageMaker, Azure, Vortex, Einstein, Watson
  - Industry / Domain AI Solution (e.g., Accenture Solutions.AI)

- **Key technical issues**
  - Input Data Quality - Sample, Shape, Coverage, Bias, Class Imbalance, Anonymity and Privacy, Noise, Drift. EDA and Feature Engineering.
  - Model Quality –
    - Traditional Metrics: Precision, Recall, f-measure, AUC
    - Benchmarks
      - Single Task (BLUE, ROUGE), Multi Task (GLUE, Super GLUE), Complex Task (BIG_Bench)
  - Inference – Representation, Realism, Trust and Explainability (local – LIME, SHAP, Counterfactuals; Global – Gradient-based), Understanding Impact

Source: Datatron
Summary

In this presentation, we discussed 5 key technology innovation trends for enterprise AI. As AI has got mainstream, it is now time to assess and analyze the medium and long-term implications of these trends, especially from the point of view of standardization, interoperability, and trustworthy adoption.
Thank you

Shubhashis Sengupta
Shubhashis.Sengupta@accenture.com
https://www.accenture.com/us-en/about/accenture-labs-index
https://www.linkedin.com/in/shubhashis/