SCALING AI AND USING IT

Bryan Catanzaro; September 9, 2021
Forward-looking statements

Except for the historical information contained herein, certain matters in this presentation including, but not limited to, statements as to: language model complexity doubling every two months; expected timing for 100T parameter single model; benefits, cost and impact of language models and few-shot learning; Megatron scaling on DGX SuperPod; reinventing NVIDIA DLSS with deep learning; AI growing at a super-exponential rate; NVIDIA building the software and systems to create the most important AI; NVIDIA building advanced AI to transform our core business; and other predictions and estimates are forward-looking statements within the meaning of the Private Securities Litigation Reform Act of 1995. These forward-looking statements and any other forward-looking statements that go beyond historical facts that are made in this presentation are subject to risks and uncertainties that may cause actual results to differ materially. Important factors that could cause actual results to differ materially include: global economic conditions; our reliance on third parties to manufacture, assemble, package and test our products; the impact of technological development and competition; development of new products and technologies or enhancements to our existing product and technologies; market acceptance of our products or our partners’ products; design, manufacturing or software defects; changes in consumer preferences and demands; changes in industry standards and interfaces; unexpected loss of performance of our products or technologies when integrated into systems and other factors. NVIDIA has based these forward-looking statements largely on its current expectations and projections about future events and trends that it believes may affect its financial condition, results of operations, business strategy, short-term and long-term business operations and objectives, and financial needs. These forward-looking statements are subject to a number of risks and uncertainties, and you should not rely upon the forward-looking statements as predictions of future events. The future events and trends discussed in this presentation may not occur and actual results could differ materially and adversely from those anticipated or implied in the forward-looking statements. Although NVIDIA believes that the expectations reflected in the forward-looking statements are reasonable, the company cannot guarantee that future results, levels of activity, performance, achievements or events and circumstances reflected in the forward-looking statements will occur. Except as required by law, NVIDIA disclaims any obligation to update these forward-looking statements to reflect future events or circumstances. For a complete discussion of factors that could materially affect our financial results and operations, please refer to the reports we file from time to time with the SEC, including our Annual Report on Form 10-K and quarterly reports on Form 10-Q. Copies of reports we file with the SEC are posted on our website and are available from NVIDIA without charge.
LARGE LANGUAGE MODELS

- Compute required to train 175B OpenAI GPT-3
  - 314 ZettaFLOP for training (3640 PFLOP/s * day)

- 100T parameter single model by 2023

Language Models Constrained by Economics
TOWARDS THE $1B MODEL
The value of few-shot learning

- These models are expensive (GPT-3 cost $12M to train)
- But their value is great
  - Generalized intelligence with few-shot learning
- What would it look like to build a $1B model?
- It would need to reinvent an entire company

English: I live in California.
Spanish: Yo vivo en California.

English: I work at NVIDIA.
Spanish: Yo trabajo en NVIDIA.

English: I believe in science.
Spanish: Yo creo en la ciencia.

Example: Few-shot translation

This model was trained on general text from the internet — yet give it a few examples and ask it to do translation, and it can translate.

Language describes all human activity.
MEGATRON SCALING ON DGX SUPERPOD

- All of NVIDIA’s HW and SW working together
- CUDA, CUDA-X AI, NVSwitch, DGX SuperPod, NCCL, CUBLAS, CUDNN
- Trained with PyTorch
- Language models useful for NVIDIA’s own products

502 petaFLOP/s sustained at 3072 GPUs
52% of tensor-core peak
163 Tflops/GPU

<table>
<thead>
<tr>
<th>Case</th>
<th>Hidden Size</th>
<th>Number of Layers</th>
<th>Model Parallel Size</th>
<th>Number of GPUs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1T</td>
<td>25600</td>
<td>128</td>
<td>512</td>
<td>3072</td>
</tr>
<tr>
<td>530B</td>
<td>20480</td>
<td>105</td>
<td>280</td>
<td>2520</td>
</tr>
<tr>
<td>145B</td>
<td>12288</td>
<td>80</td>
<td>64</td>
<td>1536</td>
</tr>
<tr>
<td>39B</td>
<td>8192</td>
<td>48</td>
<td>16</td>
<td>512</td>
</tr>
<tr>
<td>7.5B</td>
<td>4096</td>
<td>36</td>
<td>4</td>
<td>128</td>
</tr>
</tbody>
</table>

Sustained performance
Linear scaling
NVIDIA DLSS
Reinventing a Core NVIDIA Business with Deep Learning

Ray Tracing
4MPixel (1440P)
4MPixel (1440P)

Deep Learning Super Sampling
Δ W

8MPixel (4K)

VS

Supercomputer Rendered
128MPixel (16K) Ground Truth

@ctnzr
“DLSS is impressive to the point where I believe you'd be nuts not to use it.”

– Digital Foundry

“The upscaling power of this new AI driven algorithm is extremely impressive... it’s basically a free performance button.”

– Hardware Unboxed
CREATING A SELF-DRIVING CAR PLATFORM, INFRASTRUCTURE, AND SERVICE
SCALING AI AND USING IT

AI is growing at a super-exponential rate

NVIDIA is building the software and systems to create the most important AI

We are applying advanced AI to transform and grow our core business

We offer the tools for other companies to do so as well