



AI Unplugged, Modena

3 Mar 2023, Frédéric Parienté, NVIDIA



AI APPLICATION
FRAMEWORK

PLATFORMS



NVIDIA
HPC



NVIDIA
AI



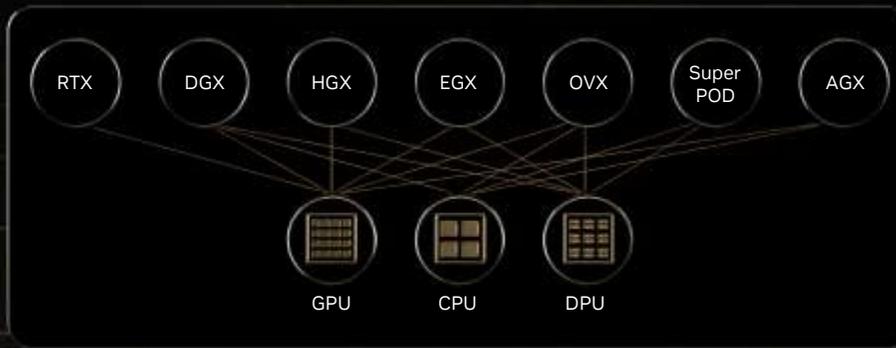
NVIDIA
Omniverse

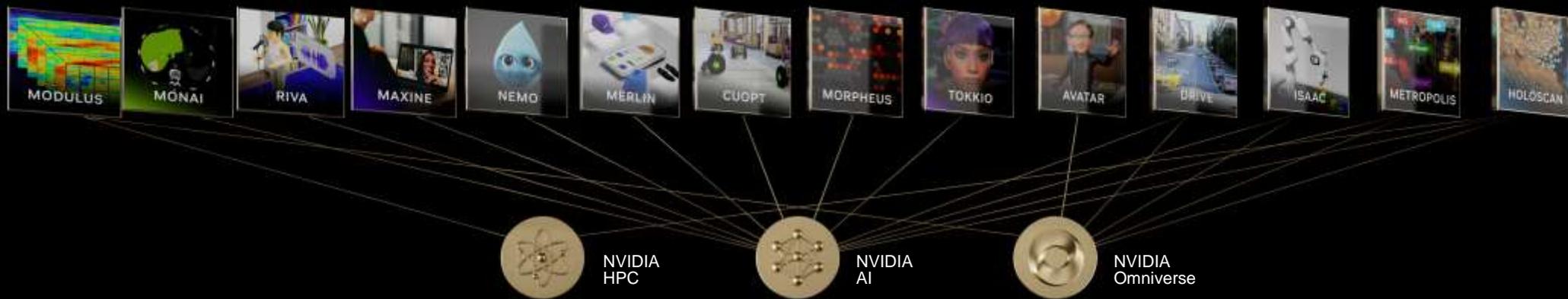
ACCELERATION
LIBRARIES



CLOUD-TO-EDGE
DATACENTER-TO-ROBOTIC SYSTEMS

3 CHIPS





3,000

Accelerated Applications

12,000

Startups

3.5M

Developers

35,000

Companies Run on NVIDIA AI

WORLD'S LEADING TECHNOLOGY PROVIDERS

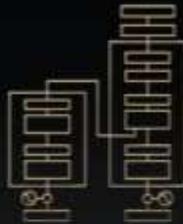


MILLION-X LEAP

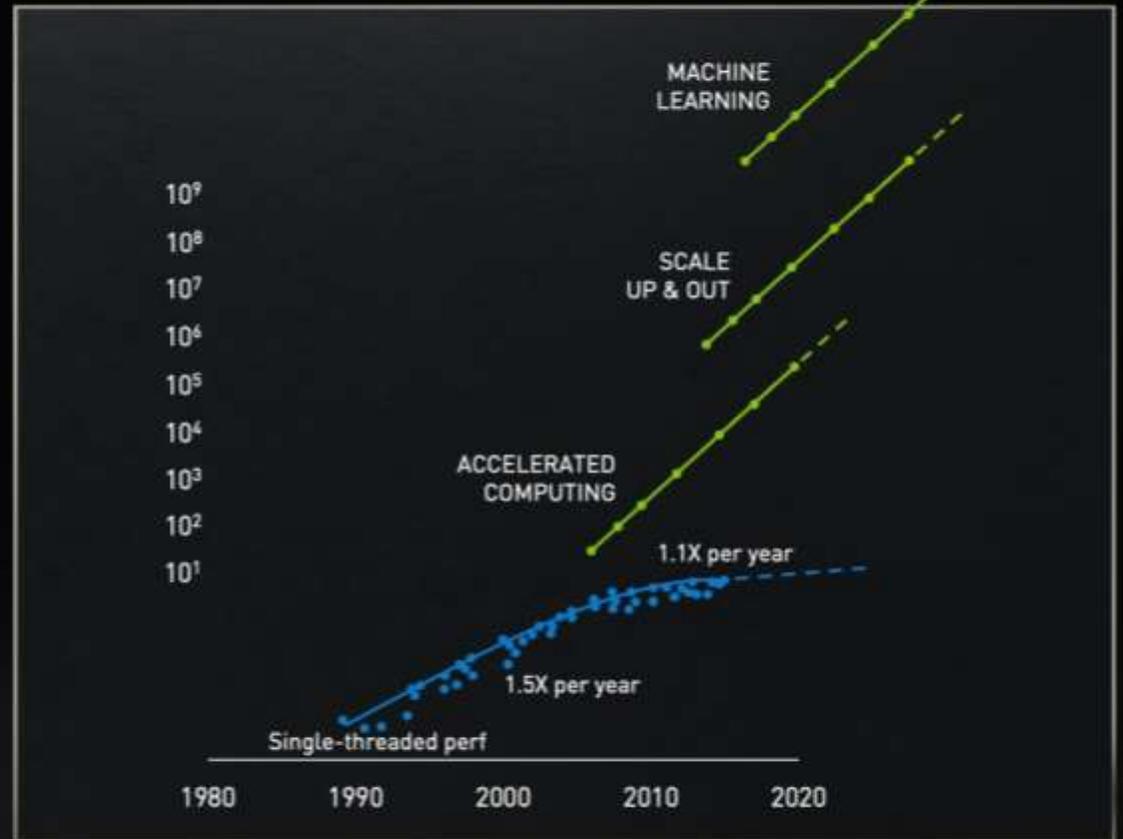
Accelerated Computing



AI



Data Center Scale





GENOMICS SIMULATION AT EXASCALE SPEED

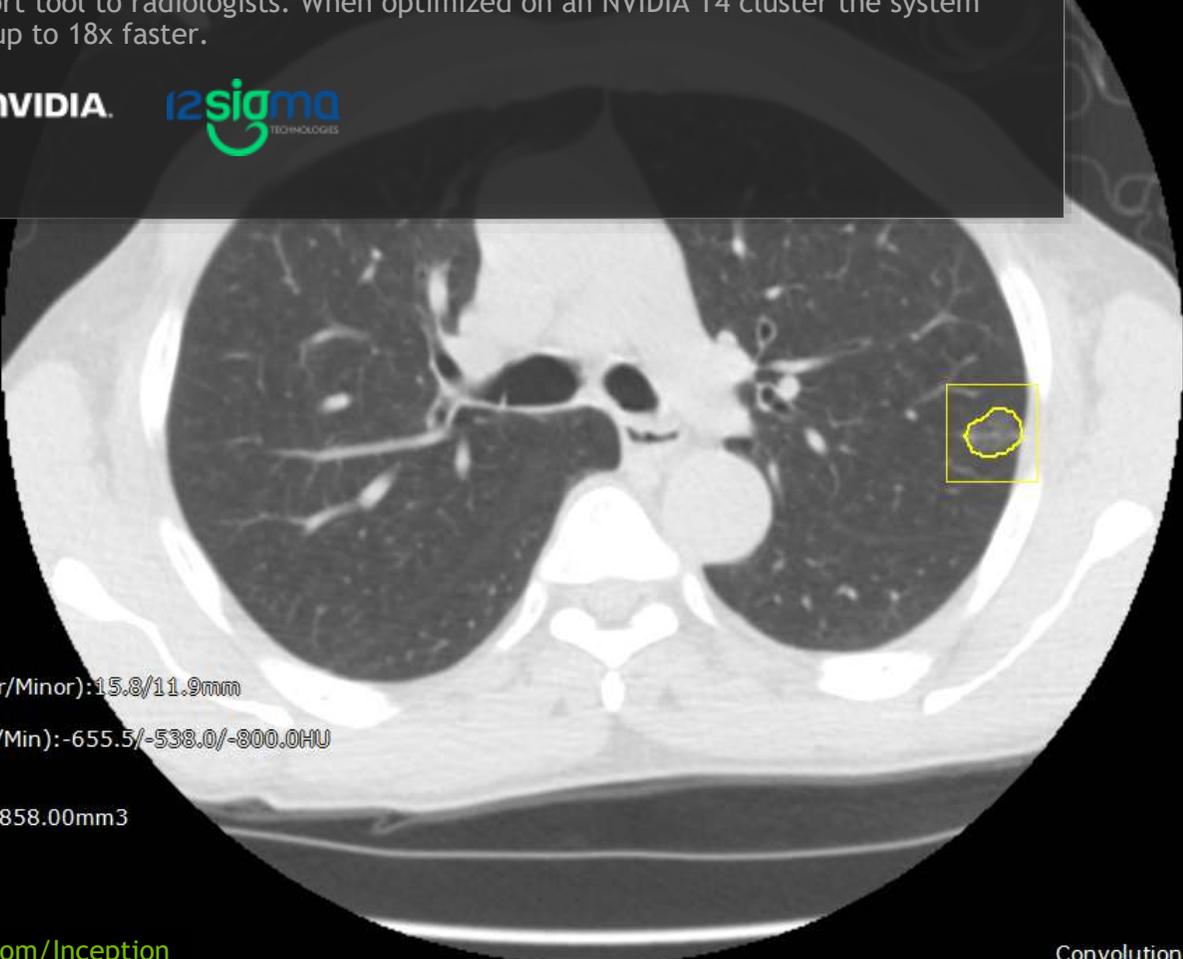
Opioid addiction was linked to ~50,000 U.S. deaths in 2017. Understanding more about how genes contribute to traits such as chronic pain and addiction could help scientists address the opioid epidemic.

A team at ORNL is comparing genetic variations to uncover hidden networks of genes that contribute to these complex traits. Using the Summit supercomputer, the team processed ~300 quadrillion element comparisons/second at a peak throughput of 2.36 exaops — the fastest science application ever reported.



USING AI TO DETECT LUNG CANCER EARLY

Powered by NVIDIA GPUs on the NVIDIA Clara platform, 12 Sigma Technologies' σ -Discover/Lung system automatically detects lung nodules as small as .01% of an image, analyzes malignancy with >90% accuracy and provides a decision support tool to radiologists. When optimized on an NVIDIA T4 cluster the system runs up to 18x faster.



3D_Diameter(Major/Minor):15.8/11.9mm
Volume:1343mm³
Pixel(Average/Max/Min):-655.5/-538.0/-800.0HU
Type:pGGO
Malignancy:90%
Lung volume:4406858.00mm³
140 kV
50 mA
848 ms

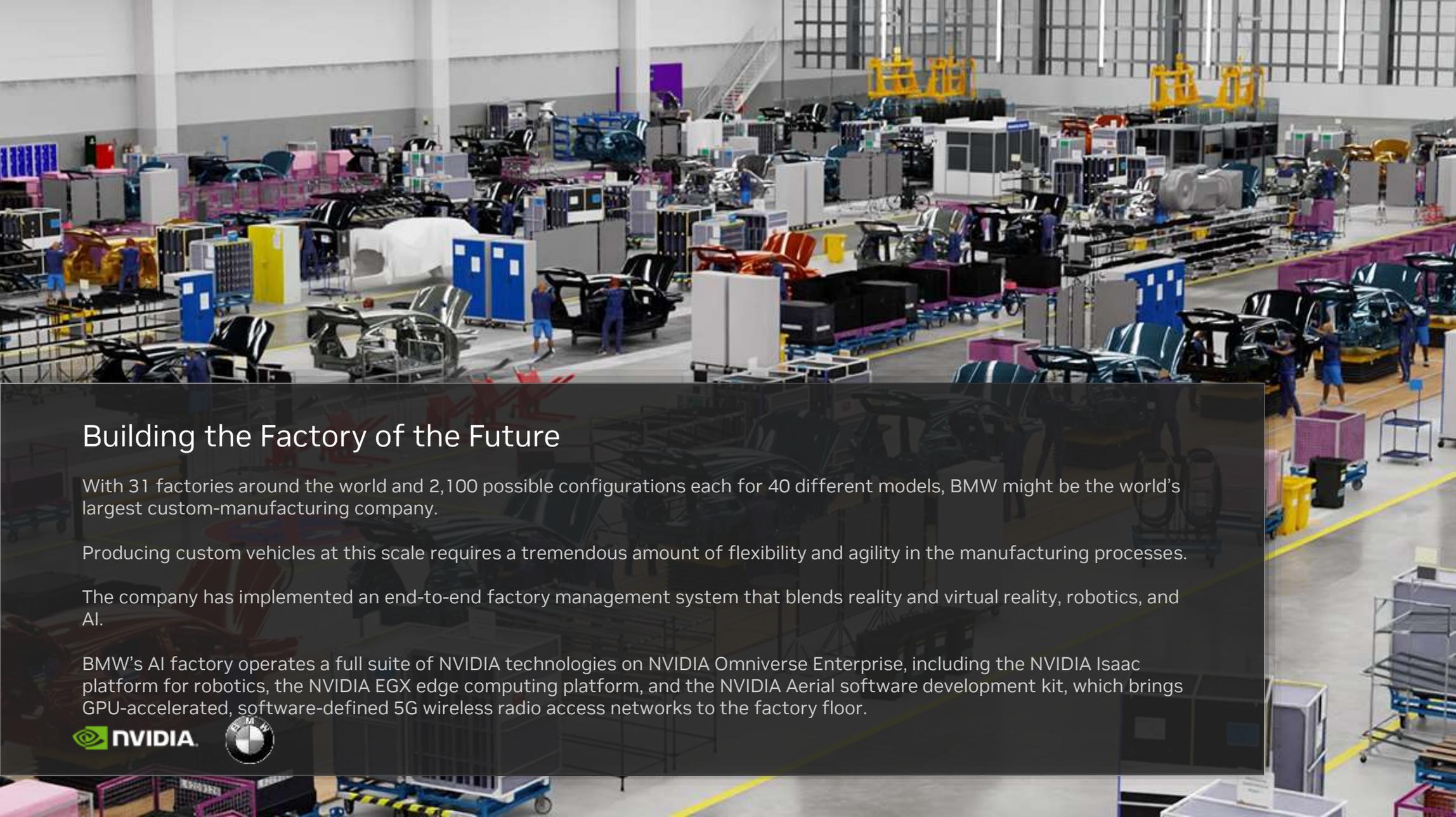
Convolution Kernel: STANDARD

www.NVIDIA.com/Inception



[Sagittal 430 / 512 »] [Average »] [Thickness 0.6 mm] + - 151977^LSS
PID: 101692 DFOV: 34.7 x 32.8 cm
Acc: 980473
1/2/2000

140 kV
50 mA
848 ms
H
R
L
F



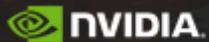
Building the Factory of the Future

With 31 factories around the world and 2,100 possible configurations each for 40 different models, BMW might be the world's largest custom-manufacturing company.

Producing custom vehicles at this scale requires a tremendous amount of flexibility and agility in the manufacturing processes.

The company has implemented an end-to-end factory management system that blends reality and virtual reality, robotics, and AI.

BMW's AI factory operates a full suite of NVIDIA technologies on NVIDIA Omniverse Enterprise, including the NVIDIA Isaac platform for robotics, the NVIDIA EGX edge computing platform, and the NVIDIA Aerial software development kit, which brings GPU-accelerated, software-defined 5G wireless radio access networks to the factory floor.





HPC+AI, OMNIVERSE PROVIDE CLEAREST VIEW OF CORONAVIRUS

A team from UCSD, the University of Pittsburg, Argonne National Laboratory and Oak Ridge National Laboratory created the first model of the coronavirus' spike protein opening and making contact with a human cell, winning a special Gordon Bell prize for HPC-based COVID-19 research.

The simulation of 305M atoms used all 27,648 NVIDIA GPUs on Summit, NVIDIA RTX 6000, NVIDIA Omniverse, NVIDIA Clara Discovery, and HPC apps that were tuned and optimized for NVIDIA GPUs including: AMBER, OpenMM, and NAMD. NVIDIA Omniverse was used to render the massive dataset in 10 days.

More than 4,000 researchers worldwide have downloaded the results — critical for vaccine design for COVID and future pathogens.



Omniverse rendering (begins at 0:52):

https://www.youtube.com/watch?v=y1IPo1KXck4&feature=emb_logo

The background features a complex pattern of thin, overlapping lines in shades of green and white against a solid black background. The lines are oriented diagonally, creating a sense of motion and depth. Some lines are sharp and bright, while others are blurred and dimmer, suggesting a light trail or a dynamic, multi-layered structure.

NVAITC Italy

From Show to Tell: A Survey on Deep Learning-based Image Captioning

Connecting Vision and Language plays an essential role in Generative Intelligence. For this reason, large research efforts have been devoted to image captioning, i.e. describing images with syntactically and semantically meaningful sentences. Starting from 2015 the task has generally been addressed with pipelines composed of a visual encoder and a language model for text generation. However, regardless of the impressive results, research in image captioning has not reached a conclusive answer yet. This work aims at providing a comprehensive overview of image captioning approaches, from visual encoding and text generation to training strategies, datasets, and evaluation metrics. In this respect, we quantitatively compare many relevant state-of-the-art approaches to identify the most impactful technical innovations in architectures and training strategies. Moreover, many variants of the problem and its open challenges are discussed. The final goal of this work is to serve as a tool for understanding the existing literature and highlighting the future directions for a research area where Computer Vision and Natural Language Processing can find an optimal synergy.

	#Params (M)	Standard Metrics						Diversity Metrics				Embedding-based Metrics			Learning-based Metrics			
		B-1	B-4	M	R	C	S	Div-1	Div-2	Vocab	%Novel	WMD	Alignment	Coverage	TIGer	BERT-S	CLIP-S	CLIP-S ^{rel}
Show and Tell ¹ [23]	13.6	72.4	31.4	25.0	53.1	97.2	18.1	0.014	0.045	635	36.1	16.5	0.199	71.7	71.8	93.4	0.697	0.762
SCST (FC) ¹ [38]	13.4	74.7	31.7	25.2	54.0	104.5	18.4	0.008	0.023	376	60.7	16.8	0.218	74.7	71.9	89.0	0.691	0.758
Show, Attend and Tell ¹ [42]	18.1	74.1	33.4	26.2	54.6	104.6	19.3	0.017	0.060	771	47.0	17.6	0.209	72.1	73.2	93.6	0.710	0.773
SCST (AH2m) ² [38]	14.5	78.0	35.3	27.1	56.7	117.4	20.5	0.010	0.031	445	64.9	18.5	0.238	76.0	73.9	88.9	0.712	0.779
Up-Down ¹ [58]	52.1	79.4	36.7	27.9	57.6	122.7	21.5	0.012	0.044	577	67.6	19.1	0.248	76.7	74.6	88.8	0.723	0.787
SGAE [71]	125.7	81.0	39.0	28.4	58.9	129.1	22.2	0.014	0.054	647	71.4	20.0	0.255	76.9	74.6	94.1	0.734	0.796
MT [72]	63.2	80.8	38.9	28.8	58.7	129.6	22.3	0.011	0.048	530	70.4	20.2	0.253	77.0	74.8	88.8	0.726	0.791
AttNet [79]	87.4	80.2	38.9	29.2	58.8	129.8	22.4	0.016	0.062	740	69.3	20.0	0.254	77.3	75.1	94.3	0.737	0.797
X-LAN [80]	75.2	80.8	39.5	29.5	59.2	132.0	23.4	0.018	0.078	858	73.9	20.6	0.261	77.9	75.4	94.3	0.746	0.803
DPA [82]	111.8	80.3	40.5	29.6	59.2	133.4	23.3	0.019	0.079	937	65.9	20.5	0.261	77.3	75.0	94.3	0.738	0.802
AutoCaption [107]	-	81.5	40.2	29.9	59.5	135.8	23.8	0.022	0.096	1064	75.8	20.9	0.262	77.7	75.4	94.3	0.752	0.808
ORT [77]	54.9	80.5	38.6	28.7	58.4	128.3	22.6	0.023	0.072	1002	73.8	19.8	0.255	76.9	75.1	94.1	0.736	0.796
CPTR [92]	138.5	81.7	40.0	29.1	59.4	129.4	-	0.014	0.068	667	75.6	20.2	0.261	77.0	74.8	94.3	0.745	0.802
M ² Transformer [81]	38.4	80.8	39.1	29.2	58.6	131.2	22.6	0.017	0.079	847	78.9	20.3	0.256	76.0	75.3	93.7	0.734	0.792
X-Transformer [80]	137.5	80.9	39.7	29.5	59.1	132.8	23.4	0.018	0.081	878	74.3	20.6	0.257	77.7	75.5	94.3	0.747	0.803
Unified VLP [101]	138.2	80.9	39.5	29.3	59.6	129.3	23.2	0.019	0.081	898	74.1	20.6	0.258	77.1	75.1	94.4	0.750	0.807
VanVL [103]	369.6	82.0	41.0	31.1	60.9	140.9	25.2	0.023	0.099	1125	77.9	20.5	0.265	79.6	75.7	88.5	0.766	0.820

COCO



Woman on a horse jumping over a pole jump.



A glass bowl contains peeled tangerines and cut strawberries.

VizWiz



A person is holding a small container of cream upside down.

TextCaps



The billboard displays 'Welcome to Yakima The Palm Springs of Washington'.

Conceptual Captions



Cars are on the streets.



Small stand of trees, just visible in the distance in the previous photo.

Fashion Captioning



A decorative leather padlock on a compact bag with croc embossed leather.

CUB-200



This bird is blue with white on its chest and has a very short beak.



UNIMORE
UNIVERSITÀ DEGLI STUDI DI MOREA



<https://arxiv.org/pdf/2107.06912.pdf>

Efficient yet Competitive Speech Translation

The primary goal of this project is to reduce model training costs without sacrificing translation quality for simultaneous speech translation tasks. As such, we first questioned the need of ASR pre-training, showing that it is not essential to achieve competitive results. Second, we focused on data filtering, showing that a simple method that looks at the ratio between source and target characters yields a quality improvement of 1 BLEU. Third, we compared different methods to reduce the detrimental effect of the audio segmentation mismatch between training data manually segmented at sentence level and inference data that is automatically segmented. Towards the same goal of training cost reduction, we participated in the simultaneous task with the same model trained for offline ST. The effectiveness of our lightweight training strategy was shown by the high score obtained on the MuST-C en-de corpus (26.7BLEU) and is confirmed in high-resource data conditions by a 1.6 BLEU improvement on the IWSLT2020 test set over last year's winning system. The result of this project got accepted for publication at IWSLT 2022.



Model	Hybrid		SHAS	
	1st-COMMON	iwslt2020	1st-COMMON	iwslt2020
1. Conformer + in-domain fn	27.4	23.8	30.3	26.4
2. Conformer_pretrain + in-domain fn	28.1	24.4	30.4	26.8
<i>with fine-tuning on resegmented data</i>				
3. Conformer + resegm. fn	28.3	25.2	29.3	26.1
4. Conformer + in-domain fn + resegm. fn	29.1	25.0	29.9	26.2
5. Conformer_pretrain + resegm. fn	29.0	25.9	29.8	26.7
6. Conformer_pretrain + in-domain fn + resegm. fn	29.0	25.7	29.7	26.8
<i>Ensembles</i>				
7. Ensemble (1, 2)	28.6	24.7	30.9	27.2
8. Ensemble (4, 6)	29.7	26.0	30.5	27.2
9. Ensemble (2, 6)	28.9	25.7	30.8	27.4
10. Ensemble (1, 2, 6)	28.9	25.8	31.3	27.6

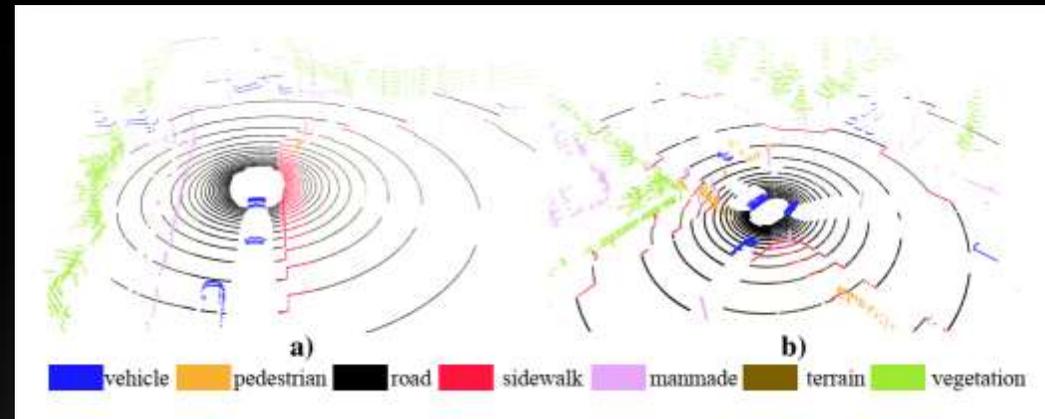
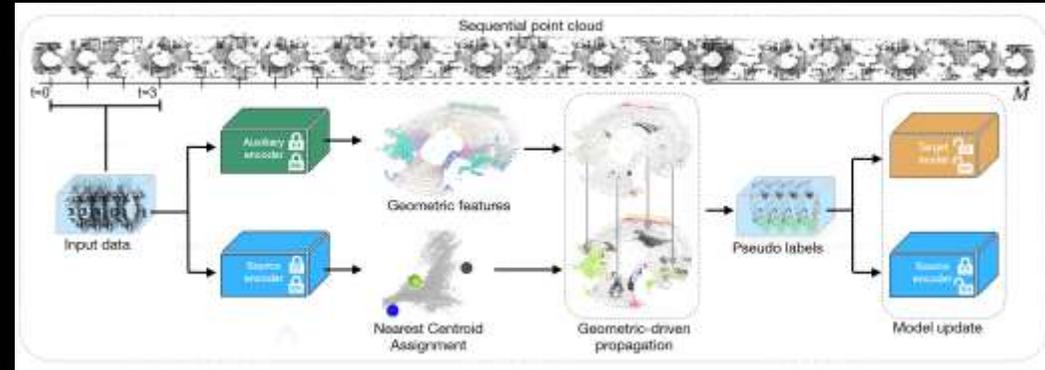
Table 4: BLEU scores of Hybrid and SHAS audio segmentation methods of the models with and without fine-tuning on re-segmented data (*resegm. fn*) on the MuST-C v2 tst-COMMON and the IWSLT2020 test set.

Model	tst2022			tst2021		
	ref2	ref1	both	ref2	ref1	both
Best direct IWSLT 2021 (Bahar et al., 2021b)	-	-	-	22.6	18.3	31.0
Best cascade IWSLT 2021 HW-TSC (Anastasopoulos et al., 2021b)	-	-	-	24.6	20.3	34.0
<i>This work</i>						
primary Ensemble (1, 2, 6)	23.6	21.0	32.9	25.5	21.3	35.6
contrastive1 Ensemble (1, 2)	23.4	20.6	32.5	25.4	20.9	35.2
contrastive2 Conformer + in-domain fn	22.8	20.1	31.6	24.5	20.2	33.9

Table 5: BLEU scores on the official blind tst2022 and tst2021 sets of our primary and contrastive submissions.

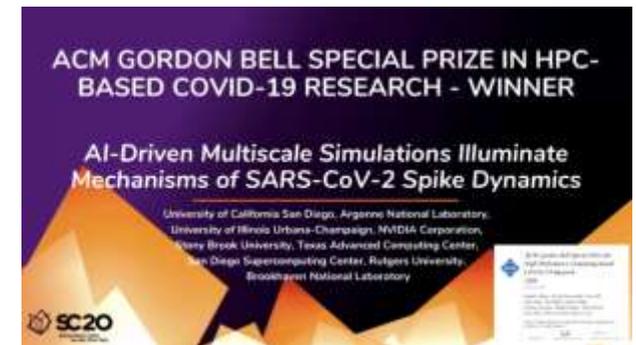
Point Cloud Domain Adaption via Forecasting

3D point cloud segmentation is a fundamental task in autonomous driving. While several approaches have been proposed in the literature, most of them neglect an important aspect, i.e. how to cope with domain shift while handling dynamic visual data. This significantly hinders the navigation capabilities of self-driving vehicles which struggle to autonomously adapt their inner models to new unseen scenarios. This work advances the state-of-the-art in this research area, by proposing the first Source-Free Online Unsupervised Domain Adaptation method for 3D LiDAR semantic segmentation, namely GP3D. Given a segmentation model trained on annotated synthetic data (source), GP3D aims to adapt this model on a target domain where no annotations are provided. Relying on feature centroids derived from the source and on geometric feature propagation, our method adapts a pre-trained source 3D segmentation model in an online fashion, without requiring neither source data nor target labels.



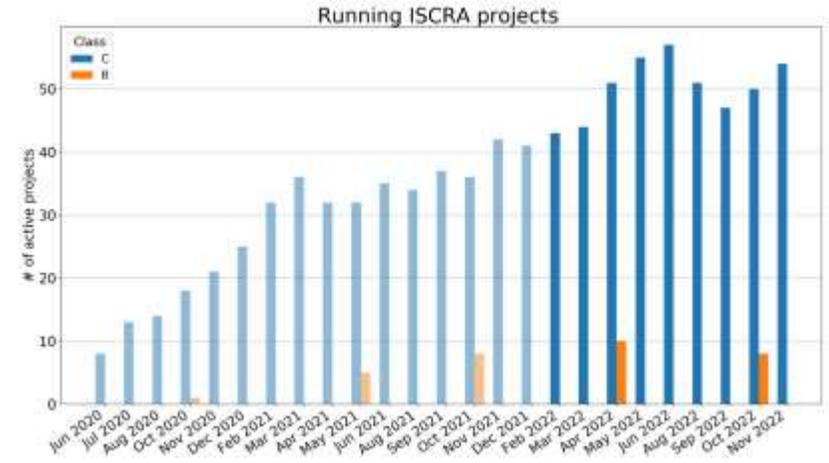
- First edition at ICPR 2020 Milan, Jan 2021 (virtual)
- Second edition at ECCV 2022 Tel Aviv, Oct 2022 (hybrid)
- Co-organized by Modena, Firenze, Berkeley, NVResearch, NVAITC
- Scientific workshop with proceedings
- ~~NVAITC meetup to boost academic collaborations~~
- RTX GPU as Best Paper award !
- Homepage cadl.it

- Submitted to ICCV 2023 Paris for a third edition 📄



Three years of activity

- **18** collaboration projects executed over various domains, from CV to astrophysics.
- More than **1300** researchers trained
- **27 joint publications** out of 52 submissions (6 still pending) – IEEE TPAMI, ECCV, ICPR, CVPR
- A **second place** at an international challenge on Action Anticipation
- **More than 50 AI specific projects enabled on CINECA (~9M core/h consumed)**
- A dedicated CINECA “Call for AI Proposals” driven by us
- Two Webinar Series in collaboration with our EMEA colleagues
- Two PhD Schools (ELLIS PhD School on Large-Scale AI - Sept 2023)
- **2 Workshops** on Computational Aspects of Deep Learning (ICPR 2020, ECCV 2022)
- **Master Degree** on Artificial Intelligence Engineering in collaboration with **University of Modena and Reggio Emilia**

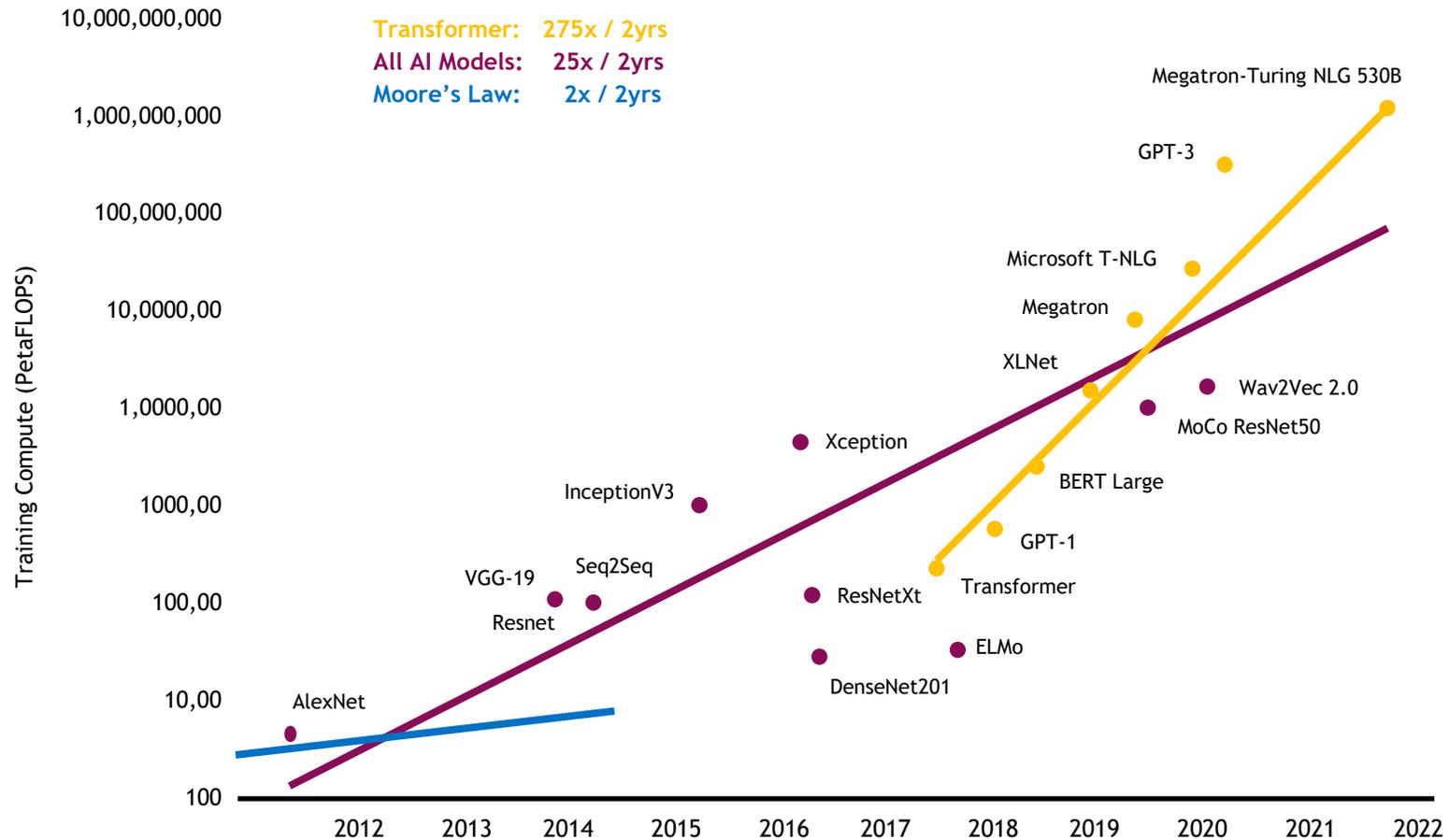


What's Next?



Dramatic increase in Model Sizes

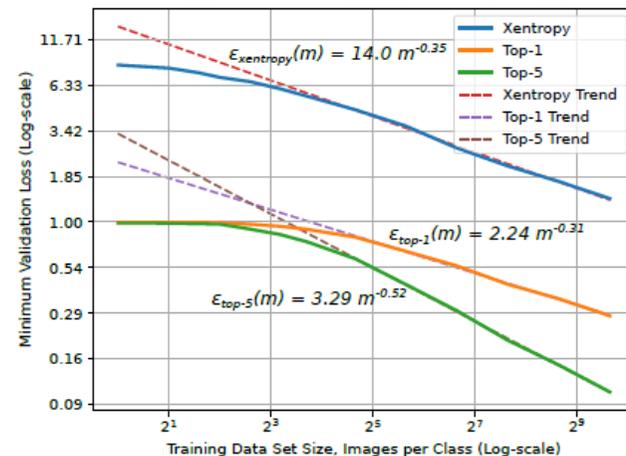
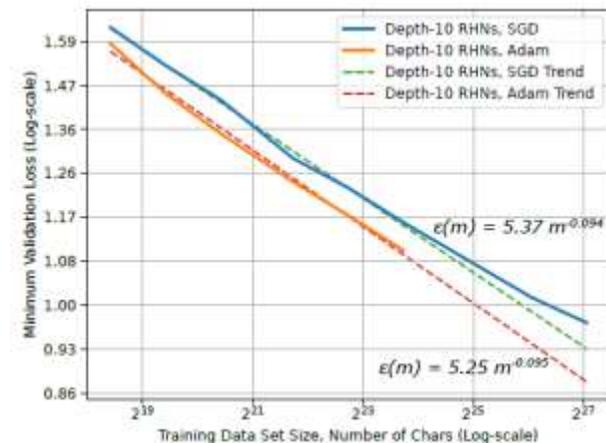
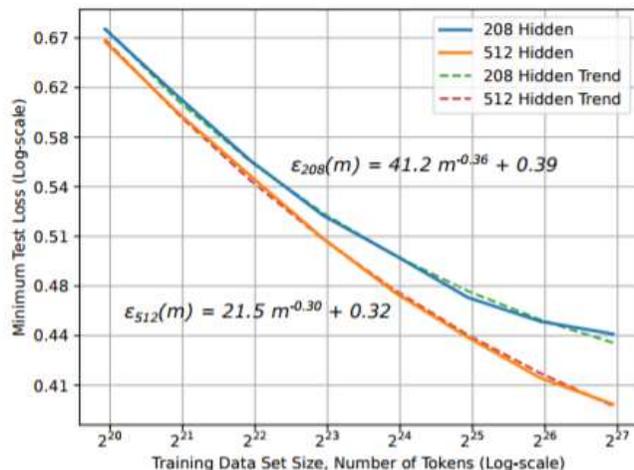
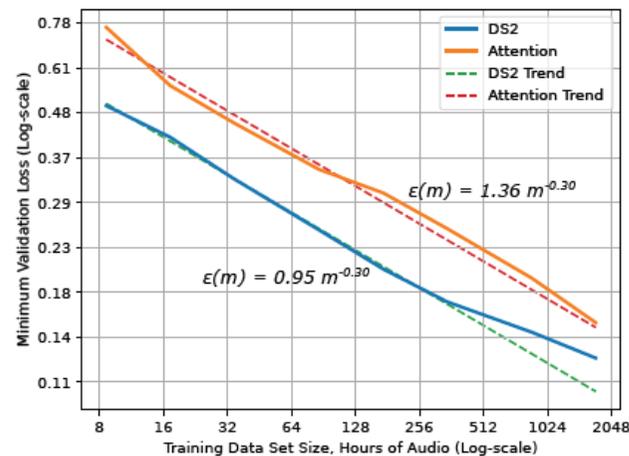
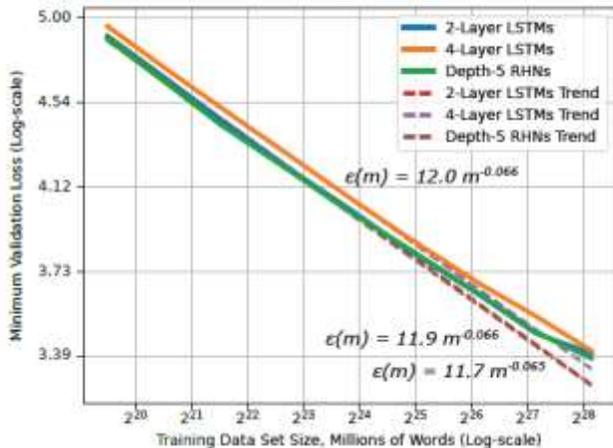
The Trend Continues



Reliable way for improving model performance

Logarithmic relationship between the dataset size and accuracy

- Translation
- Language Models
- Character Language Models
- Image Classification
- Attention Speech Models



Scaling Laws apply to NLP

As you increase the dataset size, you must also increase the model size

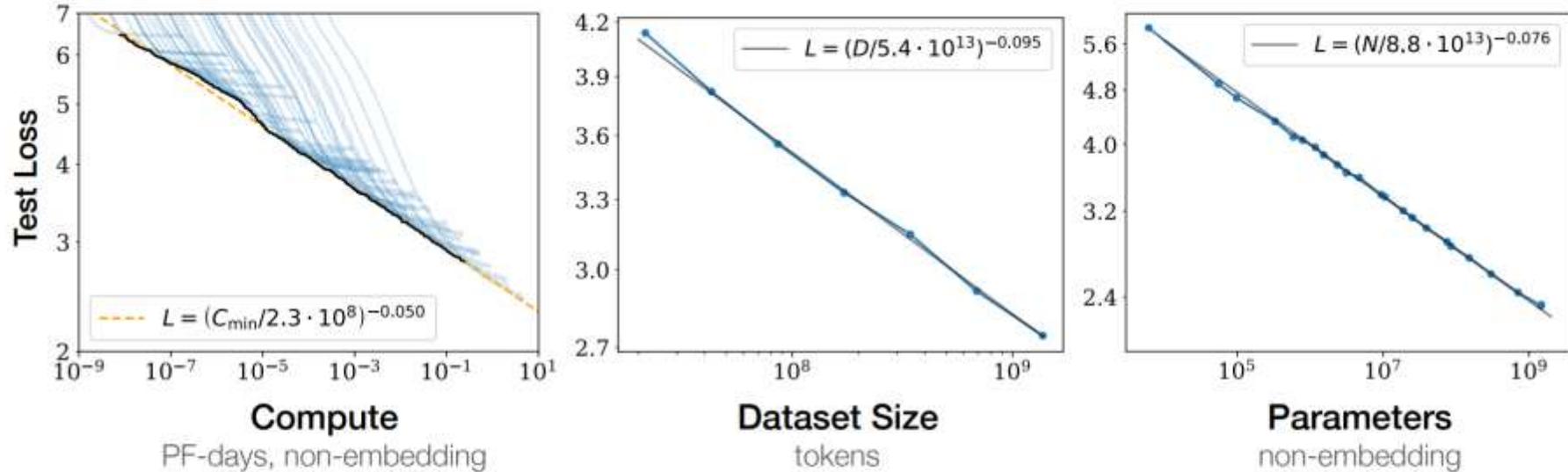


Figure 1 Language modeling performance improves smoothly as we increase the model size, dataset size, and amount of compute² used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

Scaling Laws apply to computer vision too

Increase in performance is proportional to the model size and dataset size

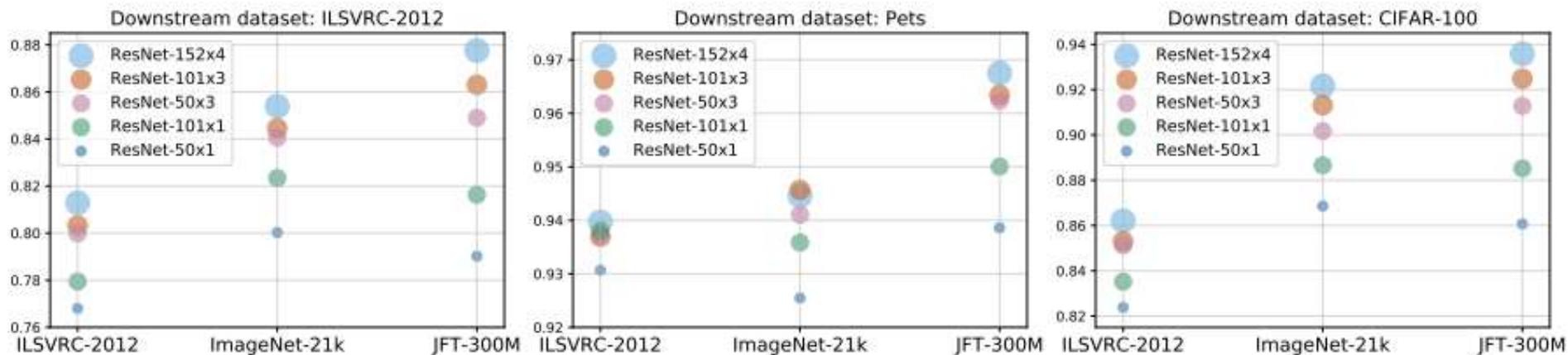
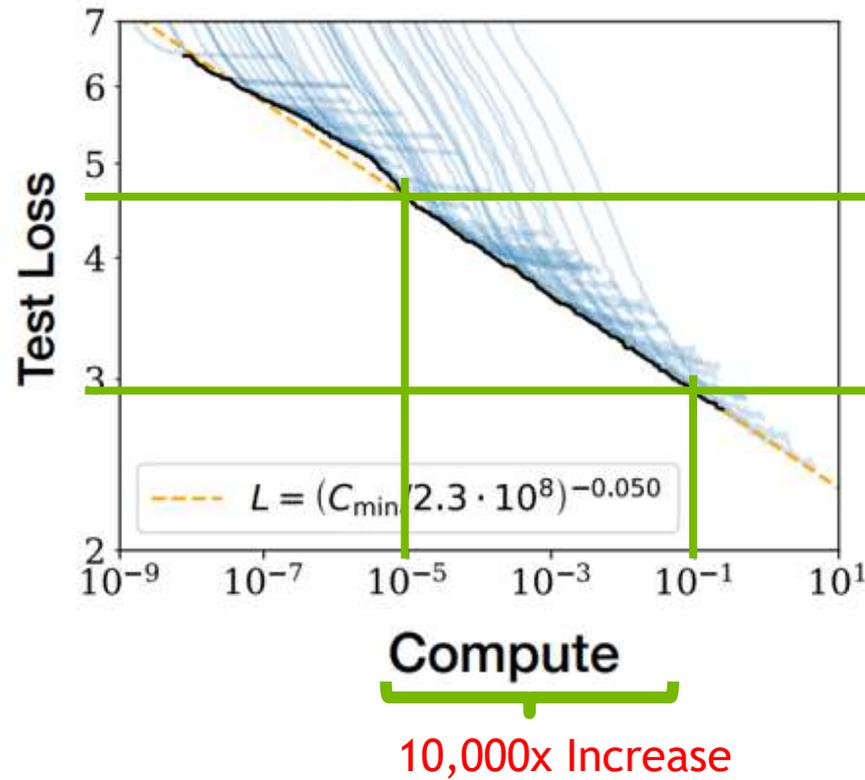


Fig. 5: Effect of upstream data (shown on the x-axis) and model size on downstream performance. Note that exclusively using more data or larger models may hurt performance; instead, both need to be increased in tandem.

Are Large language models worth it?

The cost of incremental improvement

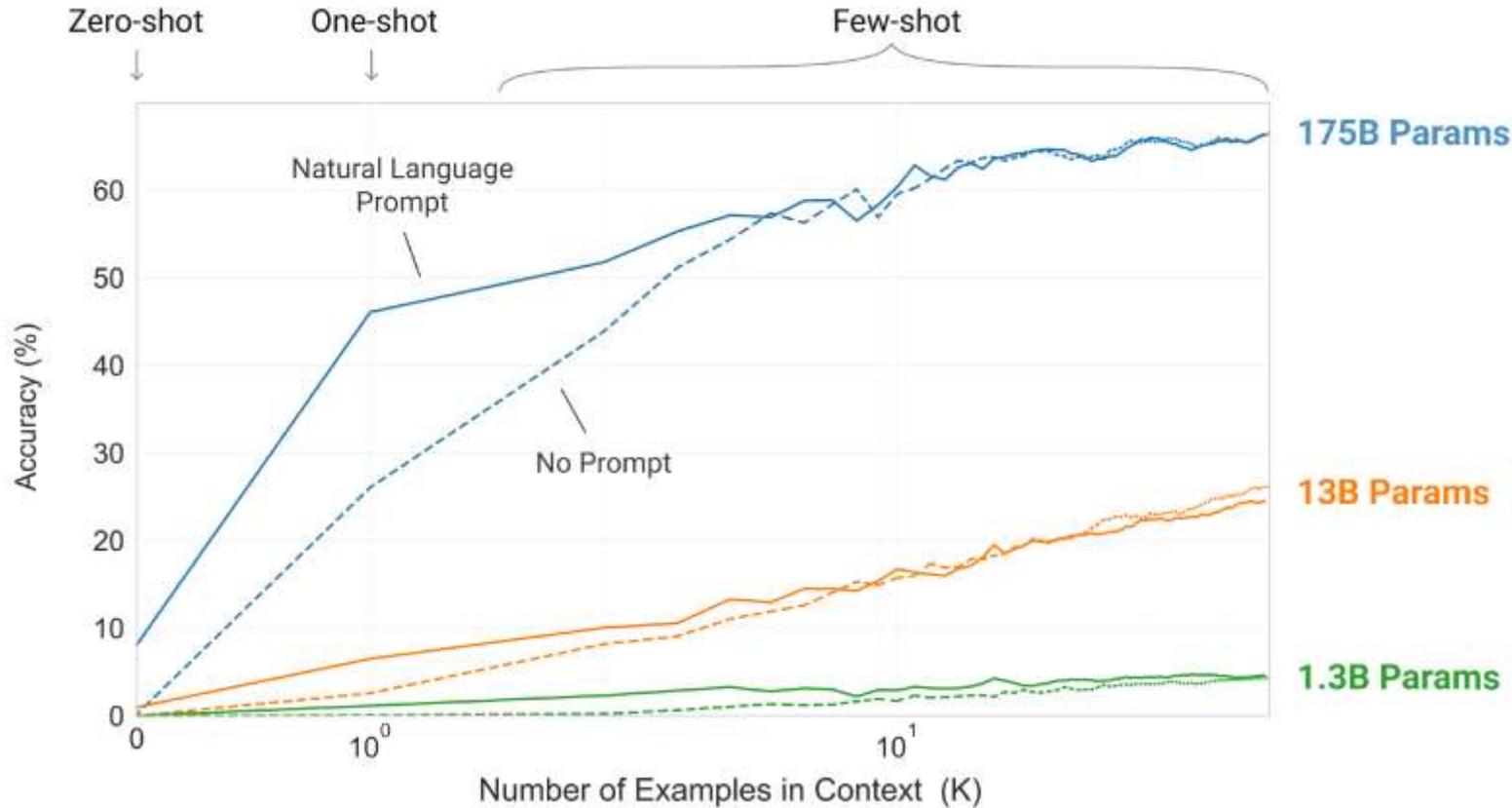


Are we building those models only for the small incremental improvement in their performance?

Is it worth all the engineering and computational investment?

Few shot learning

Learning from far fewer examples



Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1 Translate English to French: task description
2 cheese => prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1 Translate English to French: task description
2 sea otter => loutre de mer example
3 cheese => prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

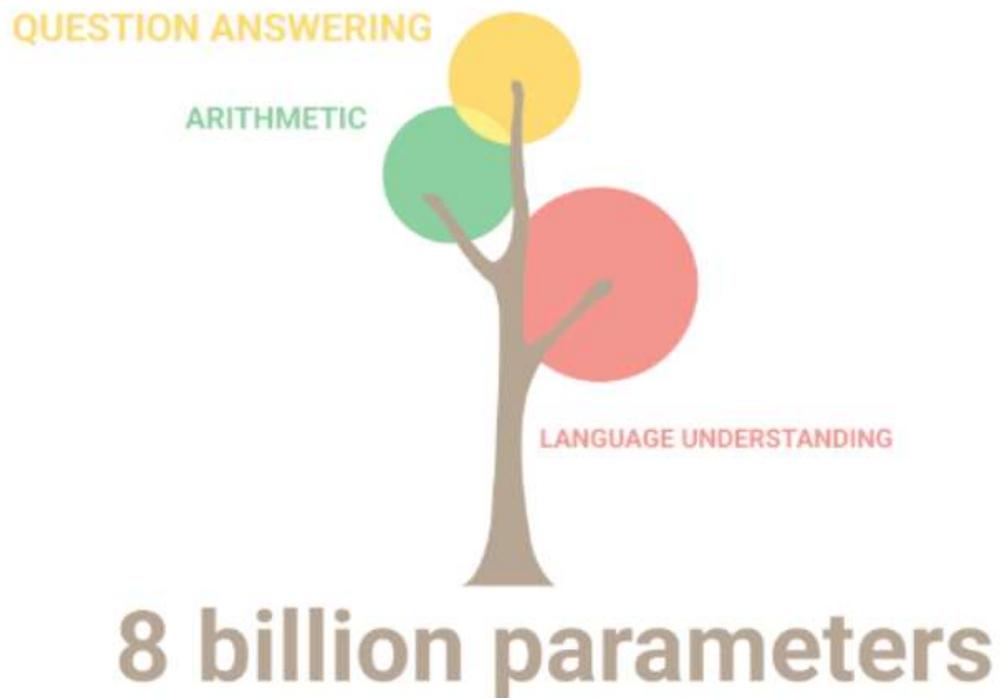
```
1 Translate English to French: task description
2 sea otter => loutre de mer examples
3 peppermint => menthe poivrée
4 plush girafe => girafe peluche
5 cheese => prompt
```

Chatbot

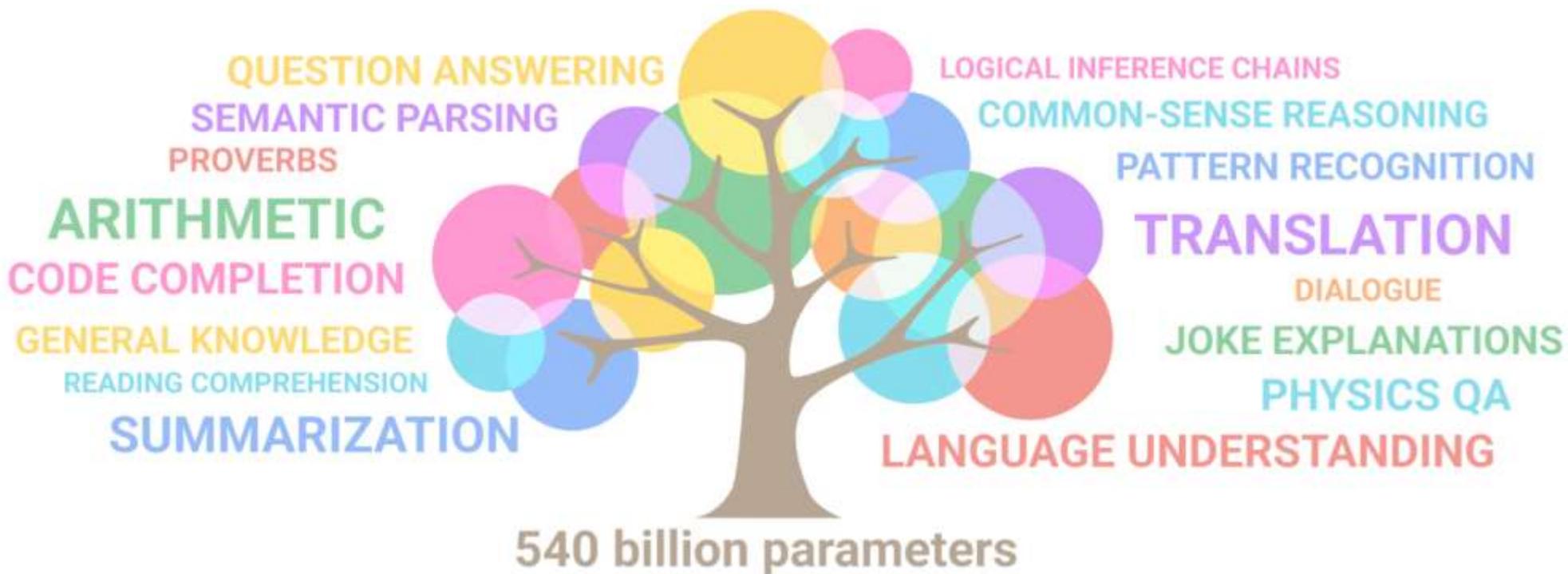
From small quantitative gains to large qualitative ones

GPT-3 1.3B param model, running on CPU	GPT-3 5B param model, running on 1 GPU	GPT-3 530B param model, running on multi-nodes
<p>Riva - 1:43 am</p> <p>Hi, welcome to Riva Q&A. How may I help you?</p> <p>You - 1:45 am</p> <p>What is the giant kelp?</p> <p>Riva - 1:45 am</p> <p>The giant kelp is a species of marine algae that grows in the temperate waters of the Southern Hemisphere.</p> <p>Riva - 1:45 am</p> <p>how does climate change affect it?</p> <p>Riva - 1:45 am</p> <p>Climate change is a process that affects the amount of heat that the oceans absorb and the amount of carbon dioxide that they release.</p> <p>Riva - 1:45 am</p> <p>where is it located?</p> <p>Riva - 1:46 am</p> <p>is found in the North Pacific Ocean.</p> <p>Riva - 1:46 am</p> <p>why is it important for the ocean?</p> <p>Riva - 1:46 am</p> <p>e ocean is the largest source of fresh water on Earth.</p> <p>Riva - 1:46 am</p> <p>Write your message... Submit</p>	<p>Riva - 1:37 am</p> <p>Hi, welcome to Riva Q&A. How may I help you?</p> <p>You - 1:37 am</p> <p>What is the giant kelp?</p> <p>Riva - 1:37 am</p> <p>Giant kelp is a brown alga.</p> <p>Riva - 1:37 am</p> <p>how does climate change affect it?</p> <p>Riva - 1:37 am</p> <p>Climate change is a major factor affecting giant kelp.</p> <p>Riva - 1:37 am</p> <p>where is it located?</p> <p>Riva - 1:38 am</p> <p>It is located in the Pacific Ocean.</p> <p>Riva - 1:38 am</p> <p>why is it important for the ocean?</p> <p>Riva - 1:38 am</p> <p>Giant kelp is an important part of the ocean food web.</p> <p>Riva - 1:38 am</p> <p>Write your message... Submit</p>	<p>You - 1:30 am</p> <p>What is the giant kelp?</p> <p>Riva - 1:30 am</p> <p>The giant kelp is a brown alga and is part of the large kingdom of life known as the Protista. Most protists are single-celled organisms, but the giant kelp is a complex species and is the largest protist in the world.</p> <p>Riva - 1:30 am</p> <p>how does climate change affect it?</p> <p>Riva - 1:30 am</p> <p>Climate change is affecting the giant kelp by warming the ocean water in which it lives. The giant kelp is sensitive to temperature changes and the warming water is causing the giant kelp to die off in large numbers.</p> <p>Riva - 1:30 am</p> <p>where is it located?</p> <p>Riva - 1:31 am</p> <p>The giant kelp is found along the Pacific coast of North America, from Alaska to Baja California.</p> <p>Riva - 1:31 am</p> <p>why is it important for the ocean?</p> <p>Riva - 1:31 am</p> <p>The giant kelp is important for the ocean because it provides food and habitat for many other species.</p> <p>Riva - 1:31 am</p> <p>Write your message... Submit</p>

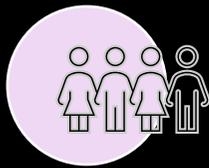
Model sizes vs tasks



Model sizes vs tasks

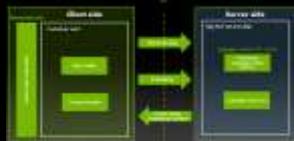


LLM JOURNEY



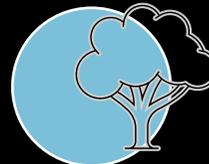
P-tuning for the masses

Bring-Your-Data and P-tune a pretrained GPT3 models to enable a variety of use cases.



Inference & Deployment

Successfully deploy 3.6B GPT3 model and make inference as trial run to pave the road for 175B GPT3 model's deployment.



Pre-training Nordic 175B GPT3

Pretraining 175B model thanks to the collaboration and the partnership.



Secure AI infra

Secure HW (Berzelius SuperPOD) and SW (NeMo Megatron) knowledge for scaling to train and deploy Nordic 175B GPT3.



More data to feed 175B

Form collaboration between Sweden, Denmark and Norway to collect more data.



Swedish GPT 3.6B model

Establish workflow within partnership and collaborations via hands-on experiments and trained the first 3.6B GPT3 model.



Start your LLM Journey

You have everything around you to succeed

- CINECA resources and support
- Strong Academy ecosystem
- Industry collaborations spurring innovation
- Large amount of material including open-source software, pre-trained models, documentation
- NVIDIA AI Technology Center (NVAITC) Italy



- Largest NVIDIA system worldwide
- 250 PFlop/s
- World's fourth most powerful supercomputer in the Top500 ranking.
- ~14K NVIDIA GPUs A100 60GB Memory each

The Conference for the Era of AI and the Metaverse

March 20-23

 NVIDIA
GTC

