The future with AI (7)

24 April 2024, 18:30 – 20:30
Cluj-Napoca, Strada Teodor Mihali 62

> From Dense Predictions to Sparse Realities:
Decoding the Future of Object Detection
Dragan Alexandru-Samuel @CTDefense

curs-ml.com
Machine Learning Course

TECH 'N TRADE

Agenda

- Introduction
- Community news
- AI News
- (Break)
- From Dense Predictions to Sparse Realities:
  Decoding the Future of Object Detection
- (Networking)

curs-ml.com
Machine Learning Course

TECH 'N TRADE

Community News

Community news

- Discord server (free to join)
  - https://discord.gg/8cG935Te

AI news

AI News

- CoreNET - Apple
- Llama 3
- Meta.AI
- Phi-3
- Synthetic data is being used all across the AI
- OpenAI and Microsoft plan a $100 billion supercomputer
- Common Corpus - 500 billion tokens of **public domain** text

CoreNET - Apple

- MLX - Apple Silicone optimized runtime for NN
    - “Apple deep learning framework similar in spirit to PyTorch, which is optimized for Apple Silicon based hardware.”

- YAML

- Object Detection Example
- Custom model Example

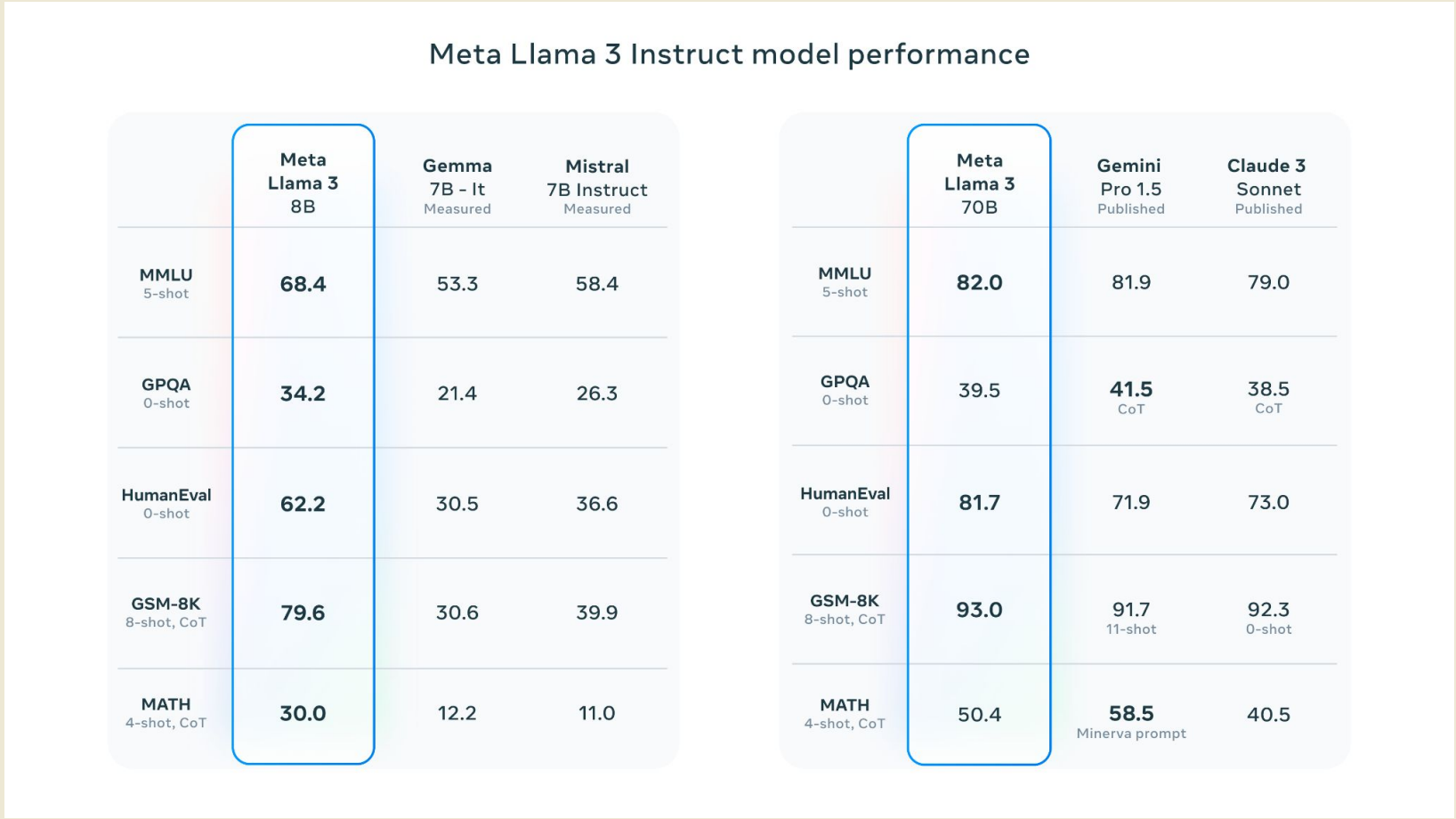
CoreNET - Apple

```python
import torch
import torch.nn.functional as F
from torch import nn

from corenet.modeling.models import MODEL_REGISTRY
from corenet.modeling.models.base_model import BaseAnyNNModel


@MODEL_REGISTRY.register("two_layer", type="classification")
class Net(BaseAnyNNModel):
    """A simple 2-layer CNN, inspired by https://pytorch.org/tutorials/beginner/blitz/cifar10_tutorial.html"""

    def __init__(self, opts: argparse.Namespace) -> None:
        super().__init__(opts)
        self.conv1 = nn.Conv2d(3, 6, 5)
        self.pool = nn.MaxPool2d(2, 2)
        self.conv2 = nn.Conv2d(6, 16, 5)
        self.fc1 = nn.Linear(16 * 5 * 5, 120)
        self.fc2 = nn.Linear(120, 84)
        self.fc3 = nn.Linear(84, 10)
        self.reset_parameters(opts)  # Initialize the weights

    def forward(self, x: torch.Tensor):
        x = self.pool(F.relu(self.conv1(x)))
        x = self.pool(F.relu(self.conv2(x)))
        x = torch.flatten(x, 1)
        x = F.relu(self.fc1(x))
        x = F.relu(self.fc2(x))
        x = self.fc3(x)
        return x
```

CoreNET - Apple

```python
from mlx_examples.open_elm import open_elm

try:
    import mlx
    from mlx import core as mx
    from mlx import nn
except ImportError:
    sys.exit("You must install Apple MLX to use this program.")


def torch_to_mlx(x: torch.Tensor) -> mx.array:
    """Converts a PyTorch tensor to an MLX tensor with the same dtype.

    Args:
        x: PyTorch tensor to convert

    Returns:
        An MLX version with the same dtype and contents.
    """
    x = x.detach()
    torch_dtype = str(x.dtype).split(".")[-1]
    mlx_dtype = getattr(mx, torch_dtype)
    # MLX mentions that converting to bfloat16 under NumPy could result in
    # precision loss, so we first up-cast to fp32.
    if torch_dtype == "bfloat16":
        x = x.to(torch.float32)
    return mx.array(x.cpu().numpy(), dtype=mlx_dtype)
```

Llama 3



Meta Llama 3 Instruct model performance

	Meta Llama 3 8B	Gemma 7B - It Measured	Mistral 7B Instruct Measured
MMLU 5-shot	**68.4**	53.3	58.4
GPQA 0-shot	**34.2**	21.4	26.3
HumanEval 0-shot	**62.2**	30.5	36.6
GSM-8K 8-shot, CoT	**79.6**	30.6	39.9
MATH 4-shot, CoT	**30.0**	12.2	11.0

	Meta Llama 3 70B	Gemini Pro 1.5 Published	Claude 3 Sonnet Published
MMLU 5-shot	**82.0**	81.9	79.0
GPQA 0-shot	39.5	**41.5** CoT	38.5 CoT
HumanEval 0-shot	**81.7**	71.9	73.0
GSM-8K 8-shot, CoT	**93.0**	91.7 11-shot	92.3 0-shot
MATH 4-shot, CoT	50.4	**58.5** Minerva prompt	40.5

Llama 3

	Training Data	Params	Context length	GQA	Token count	Knowledge cutoff
Llama 3	A new mix of publicly available online data.	8B	8k	Yes	15T+	March, 2023
		70B	8k	Yes		December, 2023

Llama 3 family of models. Token counts refer to pretraining data only. Both the 8 and 70B versions use Grouped-Query Attention (GQA) for improved inference scalability.

Llama 3

	Time (GPU hours)	Power Consumption (W)	Carbon Emitted(tCO2eq)
Llama 3 8B	1.3M	700	390
Llama 3 70B	6.4M	700	1900
Total	7.7M		2290

Llama 3

Category	Benchmark	Llama 3 8B	Llama2 7B	Llama2 13B	Llama 3 70B	Llama2 70B
General	MMLU (5-shot)	66.6	45.7	53.8	79.5	69.7
	AGIEval English (3-5 shot)	45.9	28.8	38.7	63.0	54.8
	CommonSenseQA (7-shot)	72.6	57.6	67.6	83.8	78.7
	Winogrande (5-shot)	76.1	73.3	75.4	83.1	81.8
	BIG-Bench Hard (3-shot, CoT)	61.1	38.1	47.0	81.3	65.7
	ARC-Challenge (25-shot)	78.6	53.7	67.6	93.0	85.3
Knowledge reasoning	TriviaQA-Wiki (5-shot)	78.5	72.1	79.6	89.7	87.5
Reading comprehension	SQuAD (1-shot)	76.4	72.2	72.1	85.6	82.6
	QuAC (1-shot, F1)	44.4	39.6	44.9	51.1	49.4
	BoolQ (0-shot)	75.7	65.5	66.9	79.0	73.1
	DROP (3-shot, F1)	58.4	37.9	49.8	79.7	70.2

Llama 3

Rank	Model	Arena Elo	95% CI	Votes	Organization	License	Knowledge Cutoff
1	GPT-4-Turbo-2024-04-09	1258	+4/-4	26444	OpenAI	Proprietary	2023/12
1	GPT-4-1106-preview	1253	+3/-3	68353	OpenAI	Proprietary	2023/4
1	Claude 3 Opus	1251	+3/-3	71500	Anthropic	Proprietary	2023/8
2	Gemini 1.5 Pro API-0409-Preview	1249	+4/-5	22211	Google	Proprietary	2023/11
3	GPT-4-0125-preview	1248	+2/-3	58959	OpenAI	Proprietary	2023/12
6	Meta Llama 3 70b Instruct	1213	+4/-6	15809	Meta	Llama 3 Community	2023/12
6	Bard (Gemini Pro)	1208	+7/-6	12435	Google	Proprietary	Online
7	Claude 3 Sonnet	1201	+4/-2	73414	Anthropic	Proprietary	2023/8
9	Command R+	1192	+3/-3	39716	Cohere	CC-BY-NC-4.0	2024/3
9	GPT-4-0314	1188	+3/-3	46788	OpenAI	Proprietary	2021/9
11	Claude 3 Haiku	1181	+3/-3	64518	Anthropic	Proprietary	2023/8
12	GPT-4-0613	1165	+4/-3	65523	OpenAI	Proprietary	2021/9
13	Mistral-Large-2402	1158	+3/-3	42589	Mistral	Proprietary	Unknown
13	Qwen1.5-72B-Chat	1153	+3/-3	32290	Alibaba	Qianwen LICENSE	2024/2

Llama 3 400B!

Meta Llama 3 400B+ (still training)

Checkpoint as of Apr 15, 2024

	PRE-TRAINED Meta Llama 3 400B+
MMLU 5-shot	84.8
AGIEval English 3-5-shot	69.9
BIG-Bench Hard 3-shot, CoT	85.3
ARC-Challenge 25-shot	96.0
DROP 3-shot, F1	83.5

	INSTRUCT Meta Llama 3 400B+
MMLU 5-shot	86.1
GPQA 0-shot	48.0
HumanEval 0-shot	84.1
GSM-8K 8-shot, CoT	94.1
MATH 4-shot, CoT	57.8

GPT 4 - 46.5%

Meta.AI

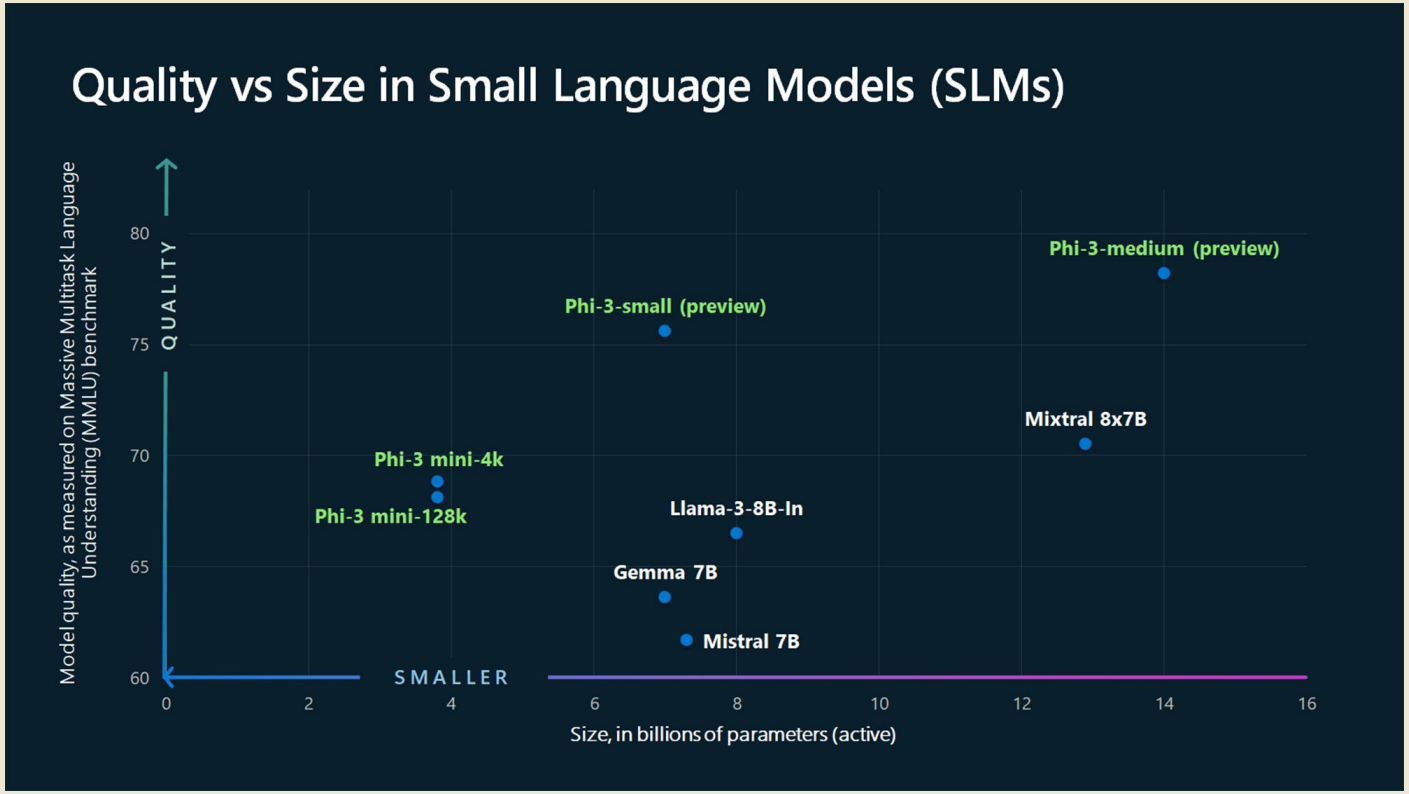
- ChatGPT equivalent from Meta
- Dialog
- Images
- Animations!
- Based on LLAMA 3 70B

- DEMO

PHI-3

- Microsoft
- 3.8B parameters
- On HF
- max. 128k tokens!
- ONNX and mobile ready

PHI-3



PHI-3

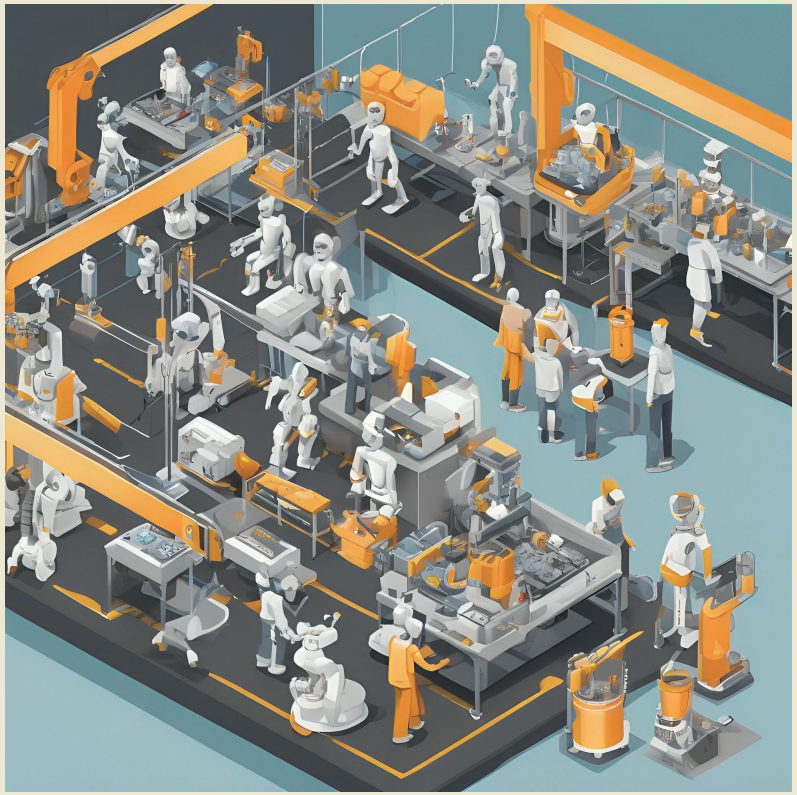
Category	Benchmark	Phi-3				Gemma-7b	Mistral-7b	Mixtral-8x7b	Llama-3-8B-In	GPT3.5-Turbo-1106	Claude-3 Sonnet
		Phi-3-Mini-4K-In	Phi-3-Mini-128K-In	Phi-3-Small (Preview)	Phi-3-Medium (Preview)						
Popular Aggregate Benchmarks	AGI Eval (0-shot)	37.5	36.9	45	48.4	42.1	35.1	45.2	42	48.4	48.4
	MMLU (5-shot)	68.8	68.1	75.6	78.2	63.6	61.7	70.5	66.5	71.4	73.9
	BigBench Hard (0-shot)	71.7	71.5	74.9	81.3	59.6	57.3	69.7	51.5	68.3	--
Language Understanding	ANLI (7-shot)	52.8	52.8	55	58.7	48.7	47.1	55.2	57.3	58.1	68.6
	HellaSwag (5-shot)	76.7	74.5	78.7	83	49.8	58.5	70.4	71.1	78.8	79.2
Reasoning	ARC Challenge (10-shot)	84.9	84	90.7	91	78.3	78.6	87.3	82.8	87.4	91.6
	ARC Easy (10-shot)	94.6	95.2	97.1	97.8	91.4	90.6	95.6	93.4	96.3	97.7
	BoolQ (0-shot)	77.6	78.7	82.9	86.6	66	72.2	76.6	80.9	79.1	87.1
	CommonsenseQA (10-shot)	80.2	78	80.3	82.6	76.2	72.6	78.1	79	79.6	82.6
	MedQA (2-shot)	53.8	55.3	58.2	69.4	49.6	50	62.2	60.5	63.4	67.9
	OpenBookQA (10-shot)	83.2	80.6	88.4	87.2	78.6	79.8	85.8	82.6	86	90.8
	PIQA (5-shot)	84.2	83.6	87.8	87.7	78.1	77.7	86	75.7	86.6	87.8
	Social IQA (5-shot)	76.6	76.1	79	80.2	65.5	74.6	75.9	73.9	68.3	80.2
	TruthfulQA (MC2) (10-shot)	65	63.2	68.7	75.7	52.1	53	60.1	63.2	67.7	77.8
	WinoGrande (5-shot)	70.8	72.5	82.5	81.4	55.6	54.2	62	65	68.8	81.4
Factual Knowledge	TriviaQA (5-shot)	64	57.1	59.1	75.6	72.3	75.2	82.2	67.7	85.8	65.7
Math	GSM8K Chain of Thought (0-shot)	82.5	83.6	88.9	90.3	59.8	46.4	64.7	77.4	78.1	79.1
Code generation	HumanEval (0-shot)	59.1	57.9	59.1	55.5	34.1	28	37.8	60.4	62.2	65.9
	MBPP (3-shot)	53.8	62.5	71.4	74.5	51.5	50.8	60.2	67.7	77.8	79.4

Synthetic data is being used all across the AI

- Google DeepMind, Stanford University, and the Georgia Institute of Technology

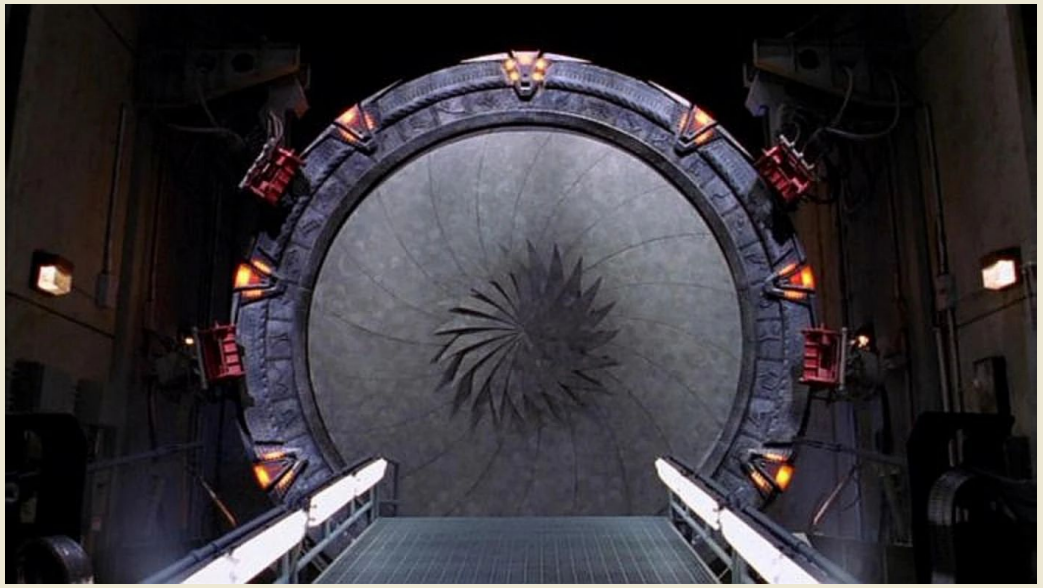
- Quality control
- Cost
- GIGO

- In:
  - training
  - evaluation
  - alignment



OpenAI and Microsoft plan a $100 billion supercomputer

- (because they can)
- named Stargate



Common Corpus - 500 billion tokens

- public domain
- available on HF
- a collection of 21 million digitized newspapers (1898 -> …)
  - English
  - French
  - German
  - Spanish
  - Dutch
  - Italian sources
  - as well as more data in other "low resource languages"

- <Right to be forgotten> ?!

(Break)

Alex Dragan is a Software Engineer recently turned Machine Learning Engineer and working to help keep you safe online. He likes challenging problems and haves an interest in learning Math and Electronics. When not working, he builds weird mechanical keyboards, plays sports/working out or enjoys video games.

From Dense Predictions to Sparse Realities: Decoding the Future of Object Detection

Thank you!

curs-ml.com
Machine Learning Course

TECH 'N TRADE