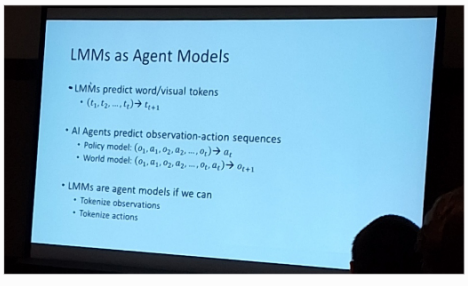
[Computer Vision in the Wild]

Opening remarks - speaker from Microsoft

- new benchmark in 2022 & 2023
- No new dataset this year

Theme: LMM in the wild

- AI agents



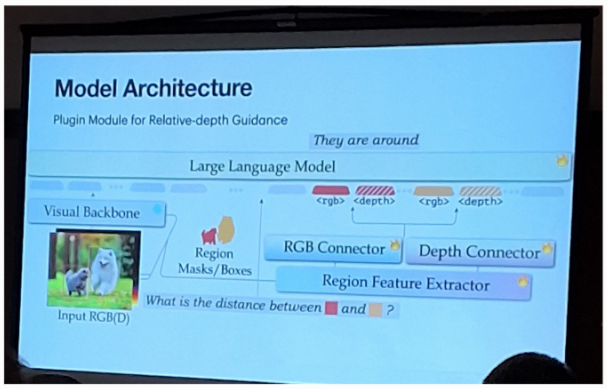
- Main point is to create specific tokenisers to accept the intended input

Invited Talk: Spatial perception & control in the wild

- Spatial RGPT

https://arxiv.org/abs/2406.01584

- Talked about dataset creation
- Able to measure distance & size in 3D space



- mentioned training only depth connector first the depth + RGD, the the rest
  - AKA freeze others

- Inject metric measurement into the training
- Someone ask how robust w/ diff camera. And the training has the camera properties estimated.

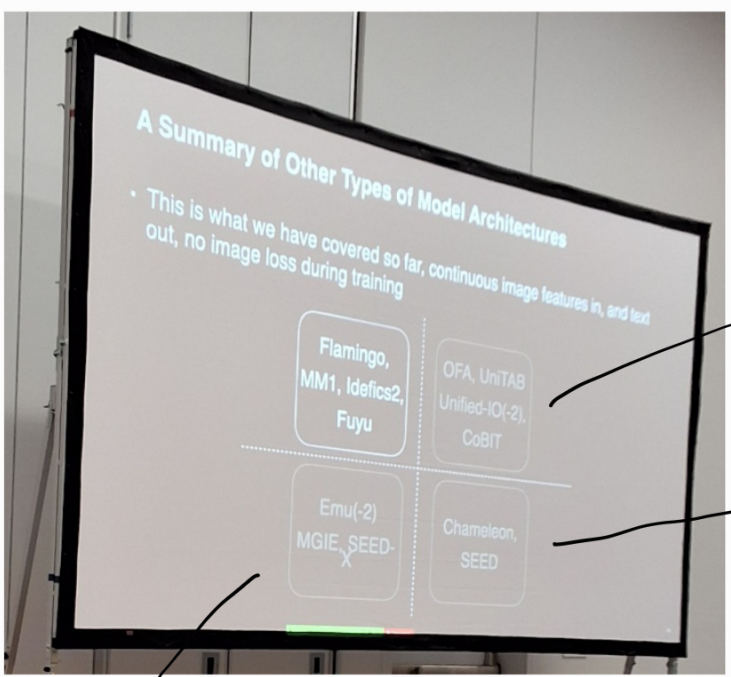
- No time for the spotlight talks
  - Blink
  - multi agent VQA
  - what's in a name
  - Vip llava

  } the repos link can be found in the webpage

[Recent advancements in vision Foundation Models]

* came in halfway

★ got recording. check webpage



(unified model example)
pro: unified archi
cons: less pop than diffusion

encoder decoder

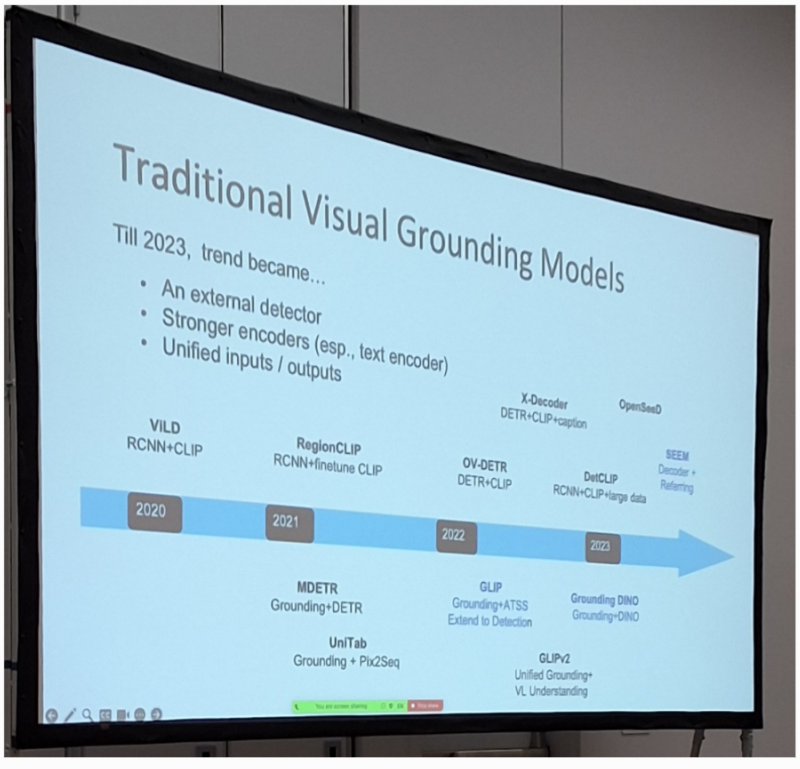
not very good multimodal benchmark

(Emu2 example)

two: strong image understanding

LLM w/ fine-grained grounding - Apple Speaker

- current MMLLM not perfect
  - ↳ 'object hallucinations'
  - ↳ not good spatial understanding
- use visual grounding to help
- visual language + grounding => better MM results
- Glip / Glip v2 < grounding Dino
  (2022) (2023)
- seem (2023)



- above is global image-perceiving, now we have fine-grained Region-level
  - ↳ training has additional tokens of region/pixels at input

https://github.com/apple/ml-ferret

  - ↳ Basically saying this is good

https://arxiv.org/abs/2404.07973v1 : Ferret v2 (no repo currently)

↳ Resolution critical to fine-grained tasks

↳ They find that CLIP encode global patch better, DinoV2 encode local patch

- Fine-grained pixel-level MLLMs
  - ↳ Vista LLM (code coming soon)
  - ↳ LISA
  - ↳ GLaMM

https://github.com/dvlab-research/LISA
https://github.com/mbzuai-oryx/groundingLMM

- video based Region-level MLLMs
  - ↳ PG-video-llava
    - ↳ also aligned audio inputs (?)

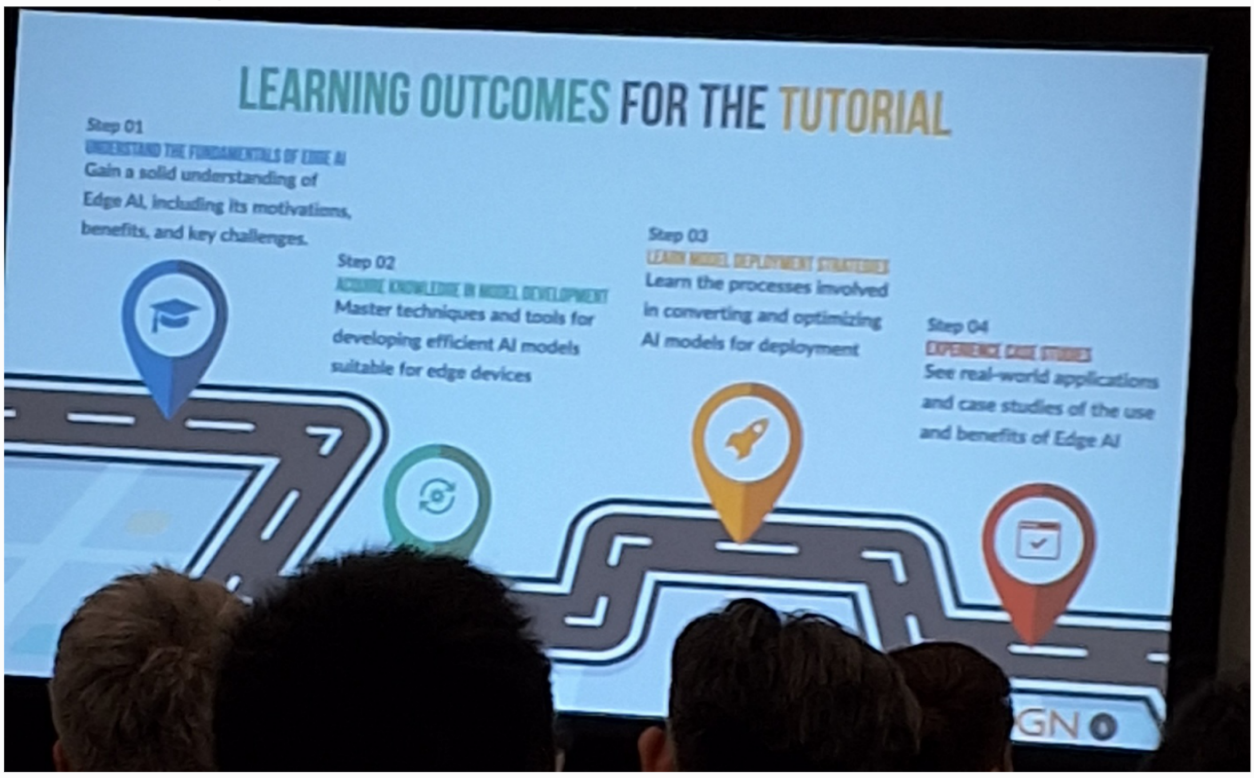
https://github.com/mbzuai-oryx/Video-LLaVA

- I think good to watch the videos (not 3D based one)

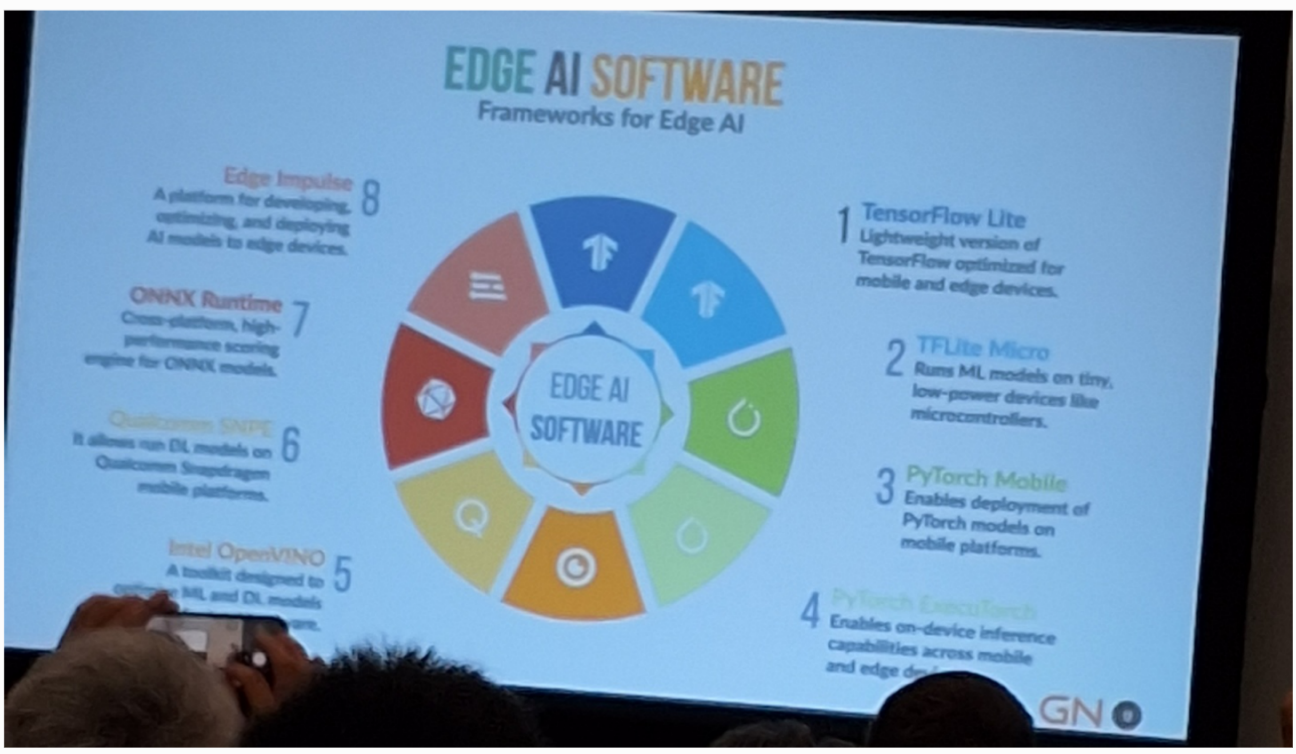
[Edge AI in artic]

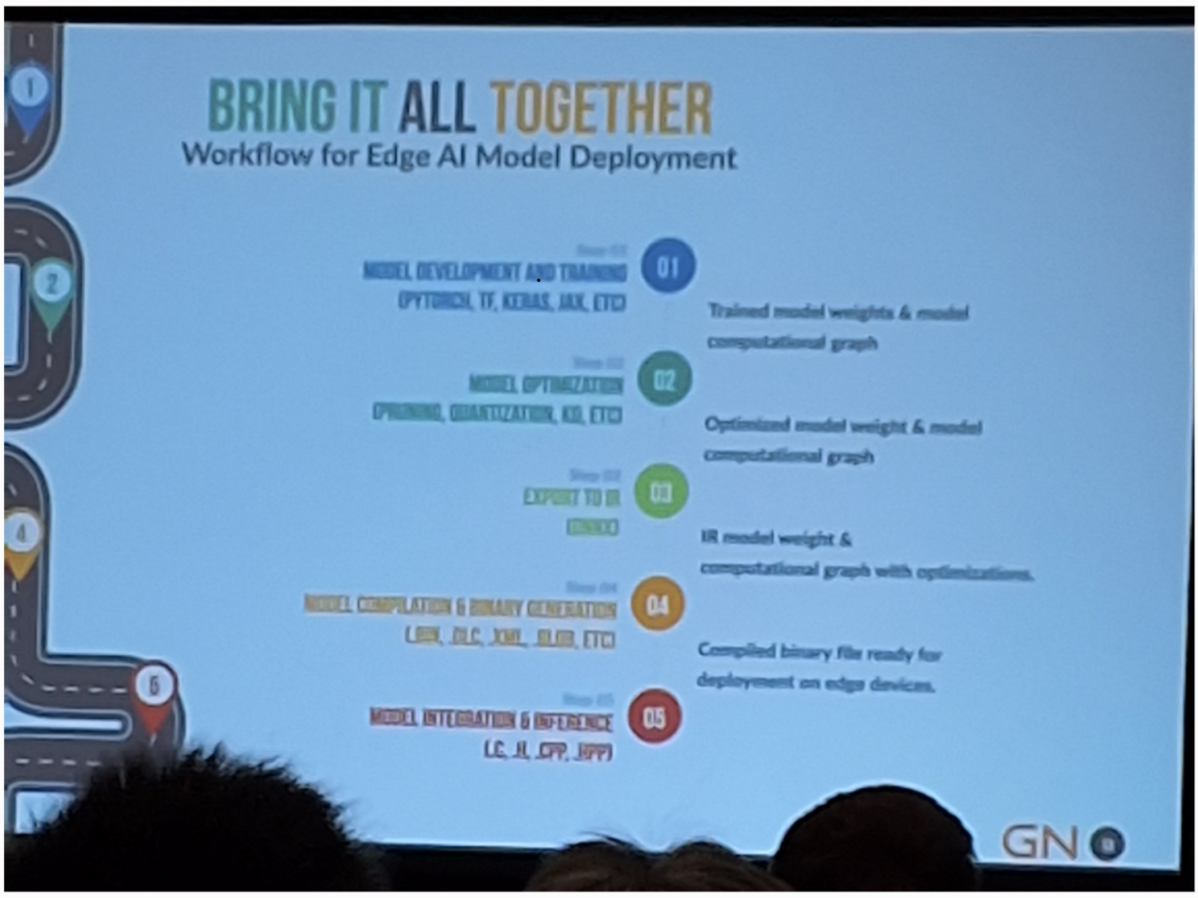
* I think I saw wrong & thought model deployment is 1pm. Actually is at 3:45 pm. Other tutorial (vision foundation + Robustness at inference) both

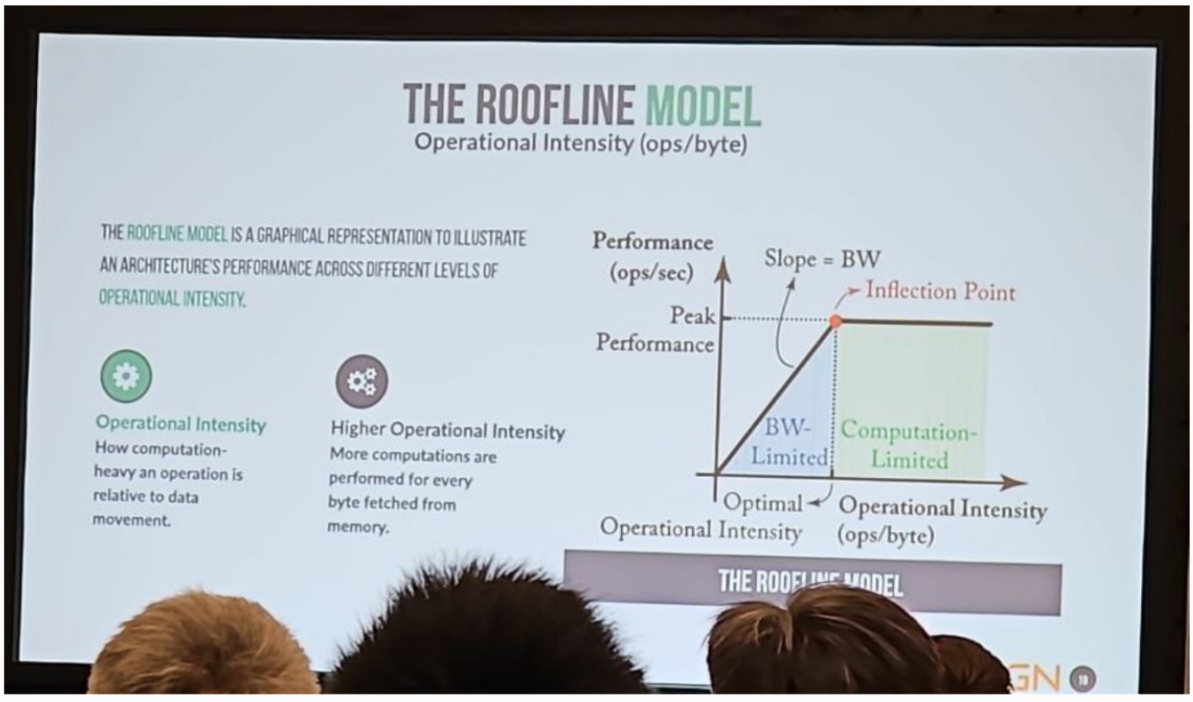
got long Q. Further, visior got record . Robust got provide slides. So I just seat in Edge AI.



Intro to Edge AI







Model development for Edge AI

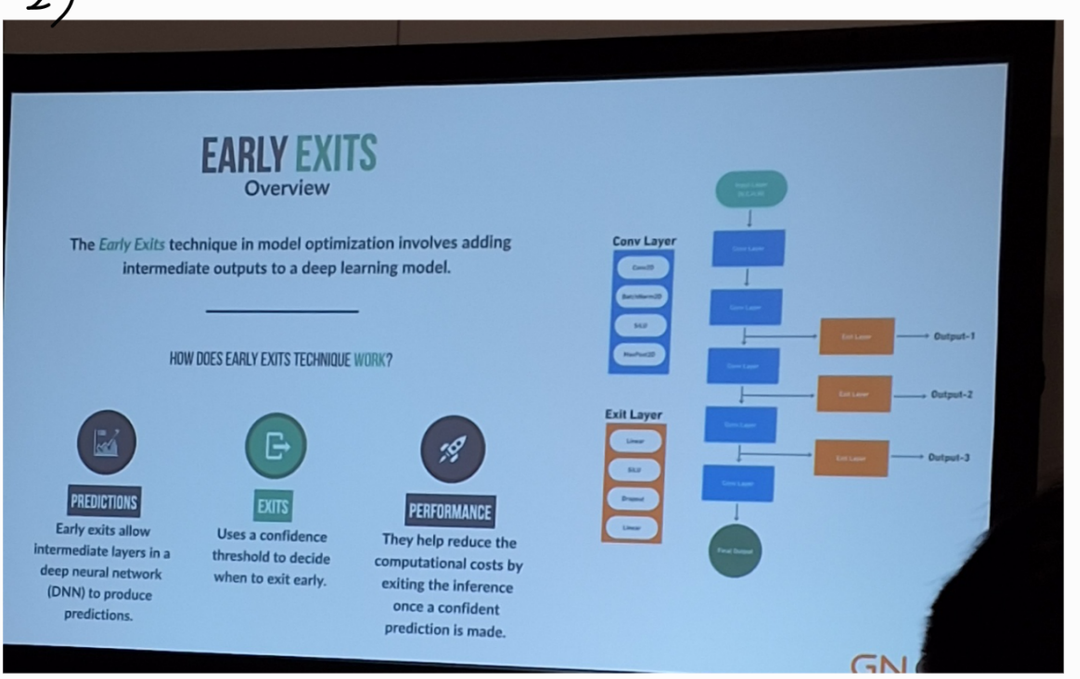
- Skip connection costly if feature size is large. Especially if is long, like from input to output
- Synthetic dataset create using blender
- 30 real, 70 synthetic gives them gd result
- Since their use case is video based, their dataset also is video based (like a video of person moving and instead of discrete images of different person at different location). This also mean that their architecture also have some temporal layers.
  - ↳ they have more video data vs static images
  - ↳ they freeze the GRU layers (temporal layer) when there is no video data in the training

Model deployment

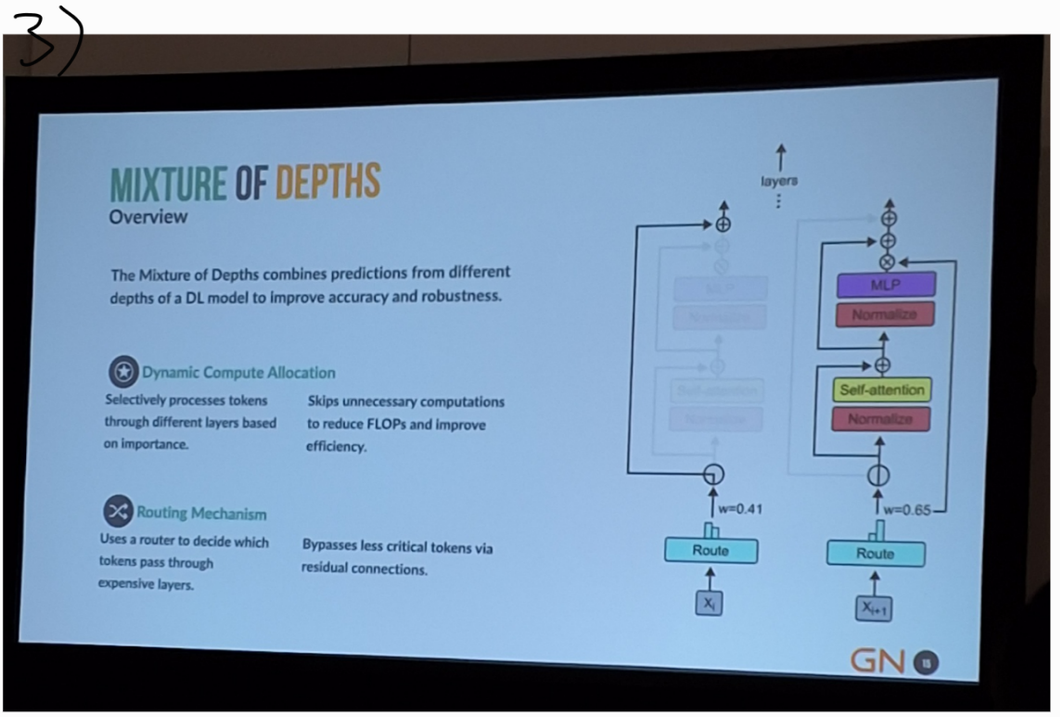
Techniques

1) Neural architecture search to find the optimal architecture first

2)

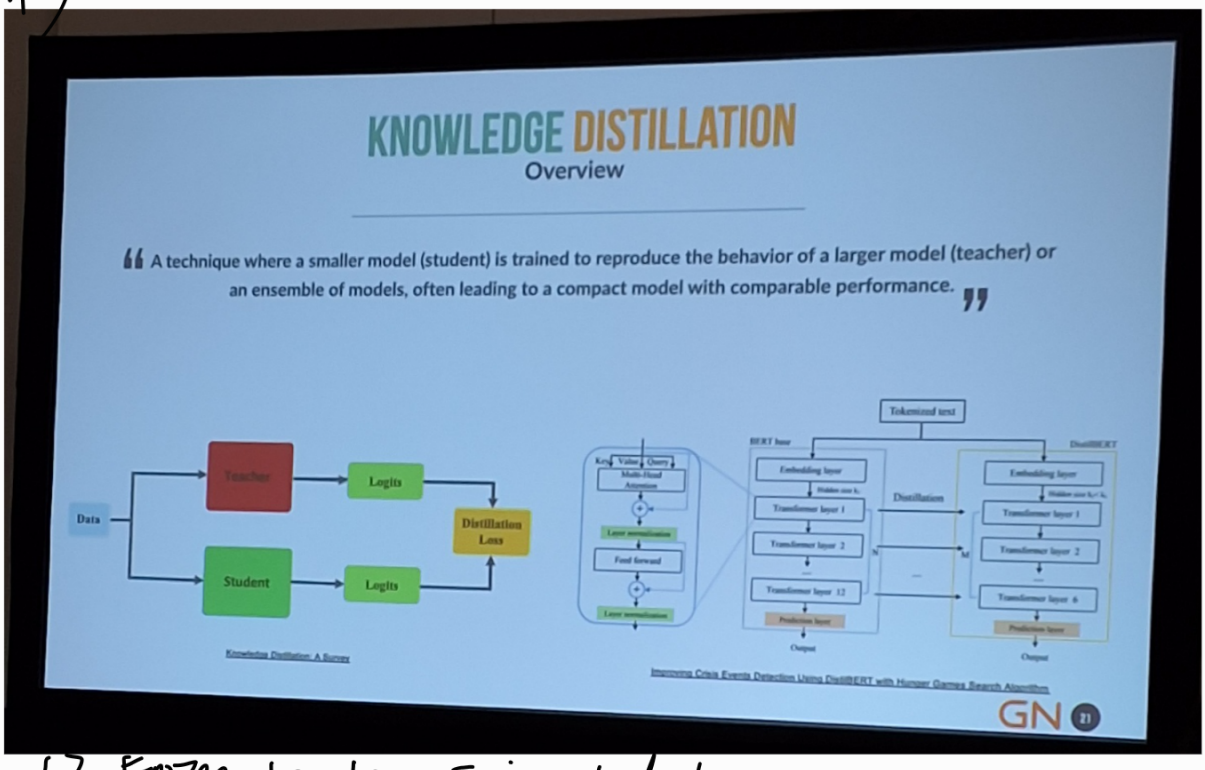


3)



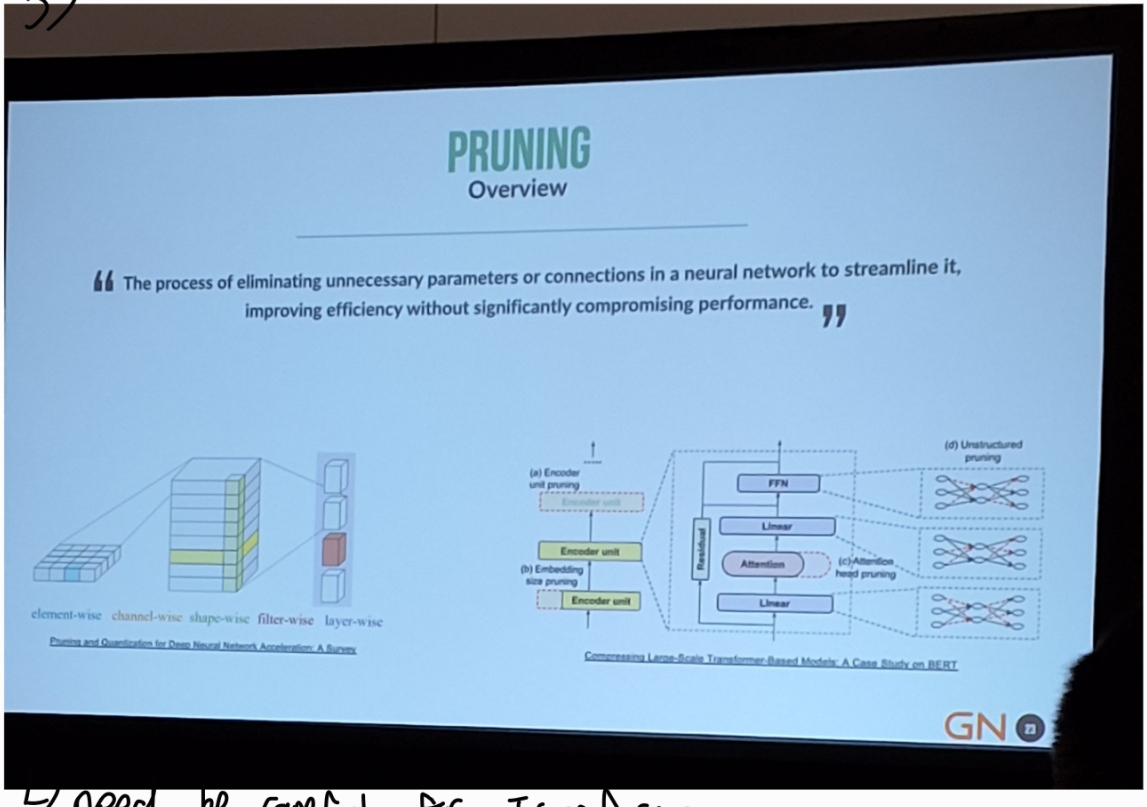
- Fast ViT

4)



↳ Frozen teacher, Train student

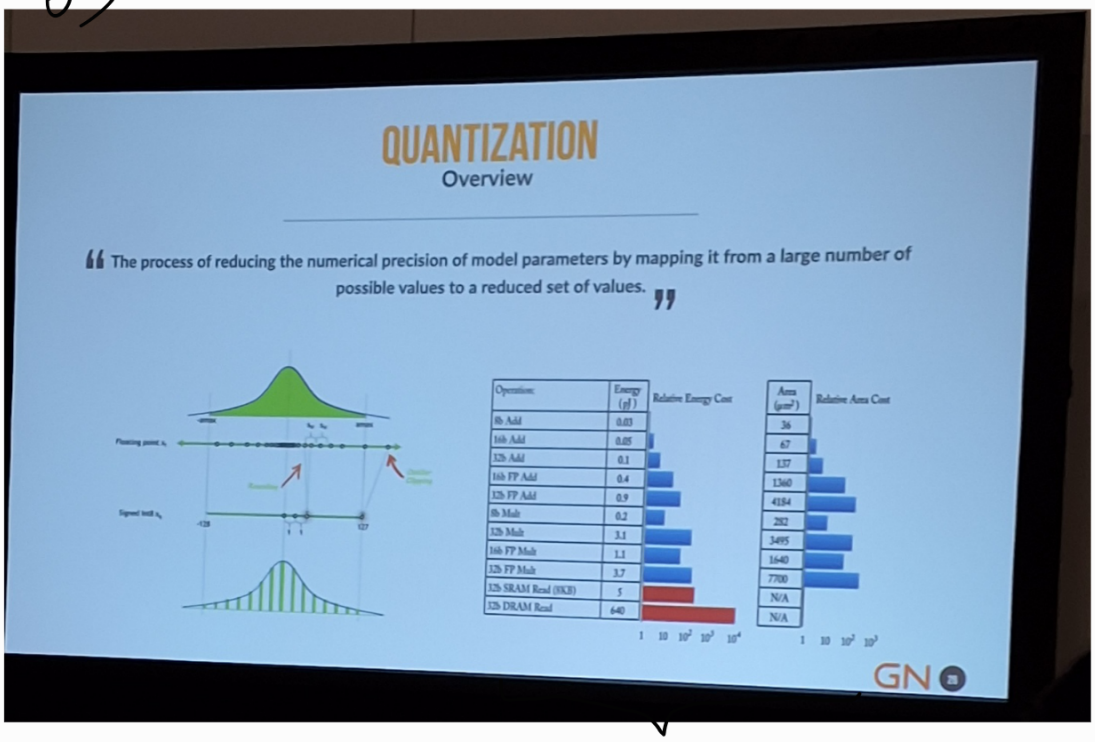
5)



↳ need be careful for Transformers

↳ only do if Hardware supports it? like some new hardware already auto only compute non-zeroes

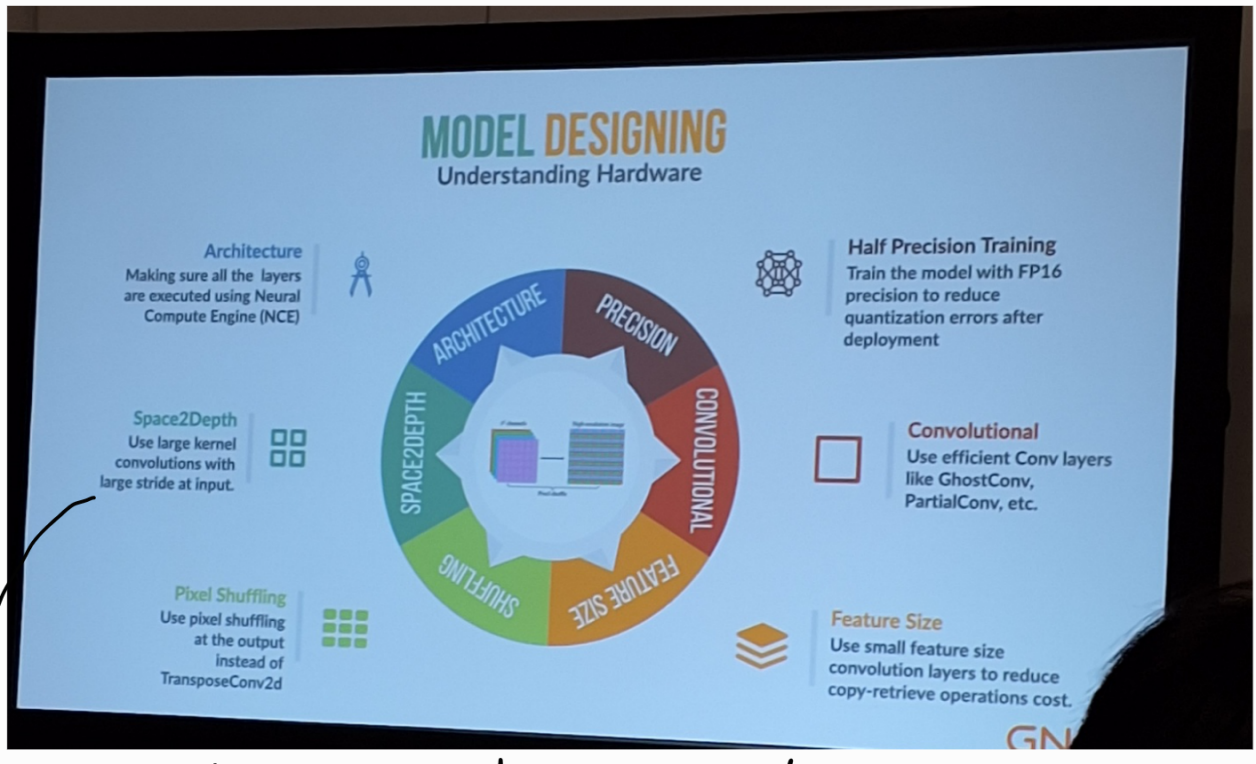
6)



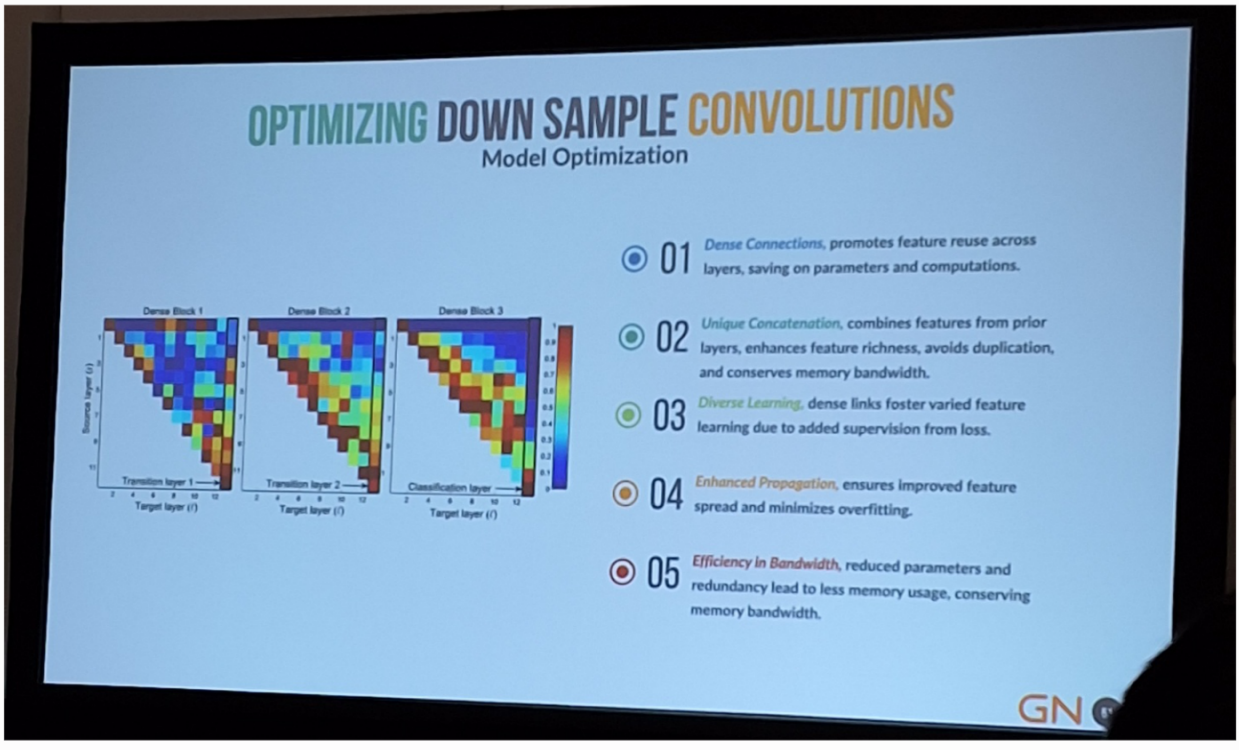
need to see hardware supports what

↳ post Training or Quantise aware training

↳ if want use, back prop shld still be higher precision else gonna → 0



↳ pushing features into depth instead of dealing with larger resolution



- Emailed computer vision in the wild & Robustness at inference
  - ↳ don't have
  - ↳ email back (Mohit). got WACV recording on Youtube