Challenges and Solutions of Moving Vision LLMs to the Edge

Costas Calamvokis

Distinguished Engineer

Expedera Inc

Presentation Introduction

- LLMs: background, underlying technologies, and growth
- How and where LLMs apply to edge AI vision
- Challenges with moving LLMs from the cloud to the edge
- What designers should consider when moving to the edge
  - The role of OEMs in facilitating Vision LLMs at the edge
- Expedera's Origin™ NPU

Costas Calamvokis
Distinguished Engineer

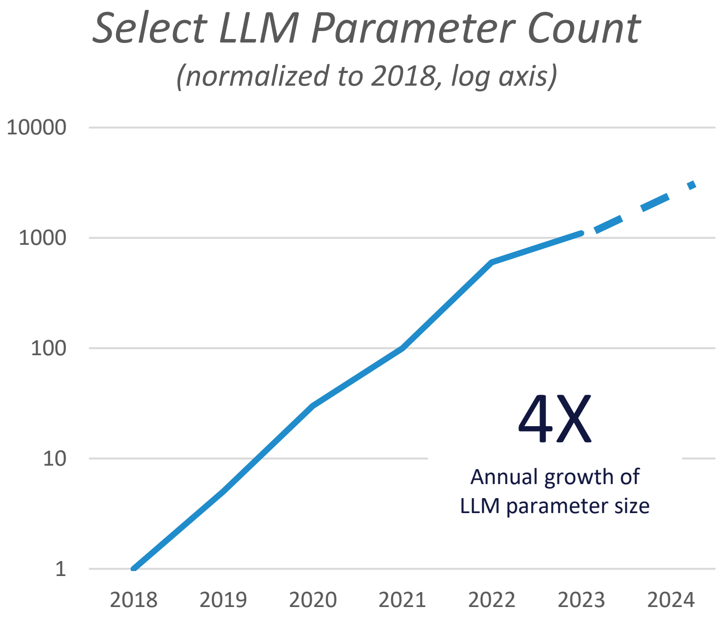
Large Language Models and Non-Language Applications

embedded VISION SUMMIT

- Large Language Models (LLMs) were designed for modeling human language
  - Language is fundamentally a structured ordering and aggregation of arbitrary objects; solutions designed for language are versatile and generalizable for many other problems
- The flexibility of LLMs in handling all kinds of data has led to the AI boom of today
  - Video, images, audio, and even computer binaries have been modeled with the tools developed for LLMs — many LLM are now multimodal: they can process different data types all in one model
  - LLMs have proven excellent at maintaining semantics, even in non-language settings

LLMs: From Large to Giant

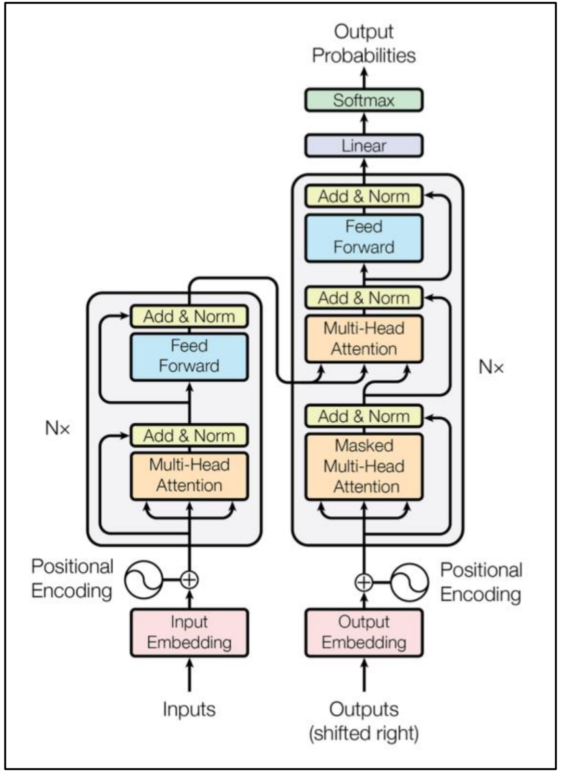
- "Large language model" can seem small by today's standards
  - Transformer (2017) maxed out at 215M parameters
  - BERT (2018) was quite large with 335M parameters
- Modern models are huge by comparison
  - GPT4 & Gemini Ultra are approximately 1.7T parameters
  - Gemini Nano-1 has 1.8B parameters
  - "Emergent" abilities such as reasoning start to appear above 1B parameters and develop most strongly towards 10B parameters and beyond.
- The largest LLMs are often cross-trained with image data



Source: McKinsey & Company 2024

Transformers: Dominating LLMs

- A strength and challenge of transformers is the attention mechanism
- Information is carried through and kept available within the context window for each token being analyzed
- All done by matrix math
- Massive weights are required, especially in the feed-forward layers and attention heads

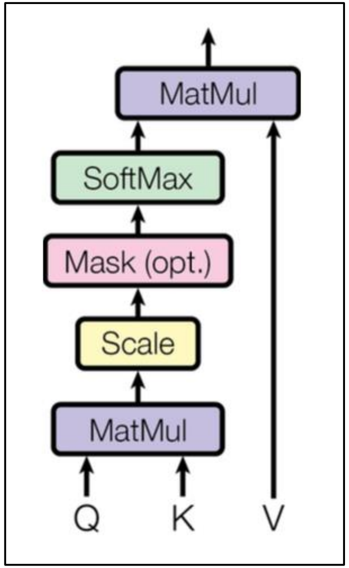


Vaswani et al 2017

Attention and the Challenge at the Edge

- Transformers are defined by their attention mechanism
  - Attention in transformers is realized as scaled dot products of the Queries (Q), Keys (K), and Values (V) matrices
- The attention mechanism is a major challenge
  - More data results in quadratic scaling of compute requirements
- Expedera's NPU has specific instructions to perform these operations with optimized data handling

$$softmax(QK^T)V$$



Vaswani et al 2017

Transformer and Non-Transformer Vision LLM Models

Transformers

- Multimodal models (Gemini, GPT4-Vision)
  - Transformer LLMs cross-trained on image data
  - The largest models allow complex discrimination of observed visual data
- Latent Diffusion Models (Stable Diffusion, Dall-E 3, Imagen)
  - U-Net model with integrated transformer modules
  - Capable of in-painting missing or obscured data as well as creative generation

Non-Transformers

- Dynalang
  - Three models, each jointly language and image trained, work together to interpret the world
  - Abstracts visual data for embodied agents
  - Allow the application of visual data to decision

Lin et al 2023

Vision LLMs on the Edge: Use Cases

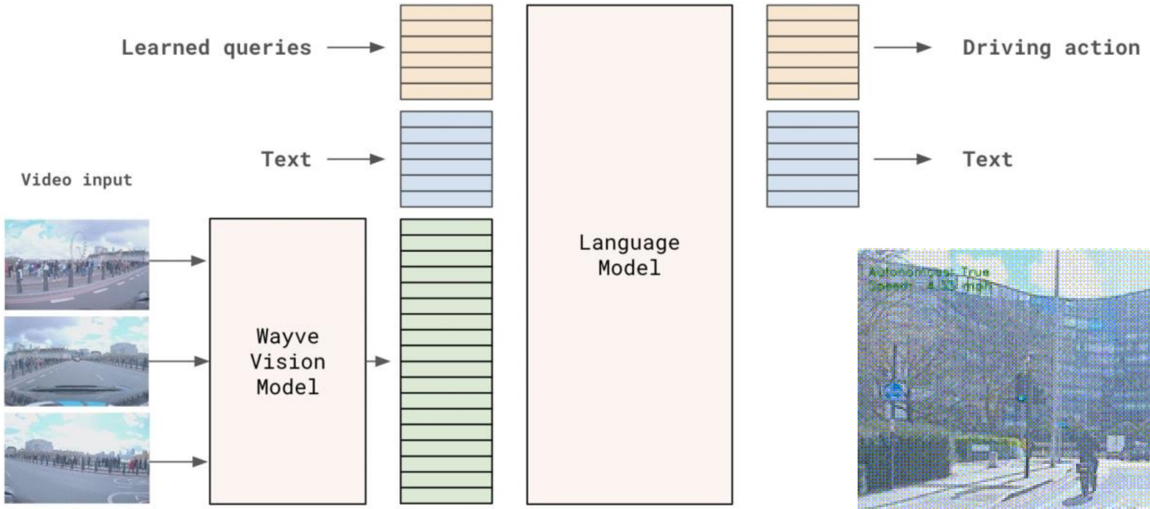
AI-Enabled Reports of Observed Events

- Video-review and analysis (Gemini — Google 2024)
- Satellite image review (CaViT — Srivastava et al 2023)
- Context-aware security: Identify violence from security footage (ViViT — Singh et al 2023)
- Driver assistance and accident prevention (LLaVA — de Zarzà et al 2023)
- Physician assistance in reviewing medical imaging (Van et al 2024; Chamblon et al 2022)

Embodied Agents

- Mobile agents, such as robots and cars, need to be able to function without a constant server link
- Language-based abstractions provide a lossy mode of "remembering" the visual inputs and reconciling them with explicit and implicit command (Dynalang — Lin et al 2023; LINGO-2 — Wayve 2024)
- Language has been demonstrated to allow the reconciling of the visually observed world with implied needs. (Dynalang — Lin et al 2023;LINGO-2 — Wayve 2024)

Use Cases: LINGO-2 & Language-Directed Driving



Wayve 2024

Design Challenges of LLMs on the Edge

- LLM models are compute- and memory-intensive
  - Increased parameters = increased data and processing requirements
- LLMs have been mostly cloud-centric
  - Adequate processing and no major power issues, but with concerns about latency and privacy in mission-critical use cases
- Even ‘modest’ all-language 7B parameter models have struggled to run on edge hardware at user-friendly rates
  - Vision LLMs will need to be fast with minimum latencies to meet use case requirements

Optimizing Vision LLM Edge Deployments…

Model Architectures

- Alternative architectures, such as Hungry Hungry Hippo (H3) modules replacing transformer blocks
- Changing how and where transformer modules are used (e.g., SDXL)
- "Distilled" models

System Resource Utilization

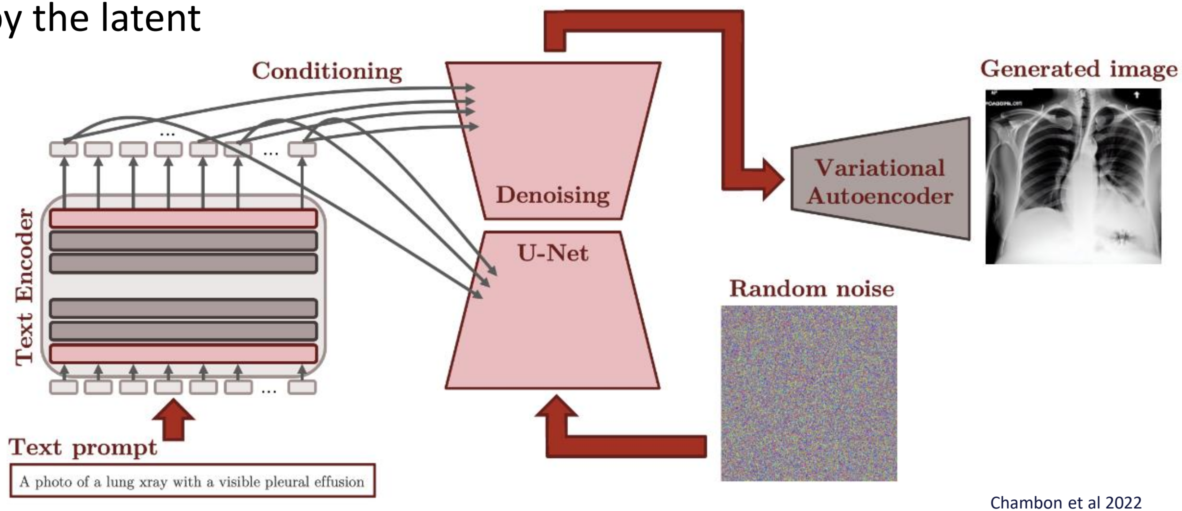
- Quantization reduces compute complexity and memory demands at the cost of accuracy
- Tiling, such as FlashAttention, improves how models are fed through the processors
- Speculative decoding can pre-guess pending results

Dedicated Hardware Support

- Standard vs bespoke processors
- General support vs tailored to specific use cases
- Trade-off between versatility vs utilization, throughput, power consumption, silicon footprint differences

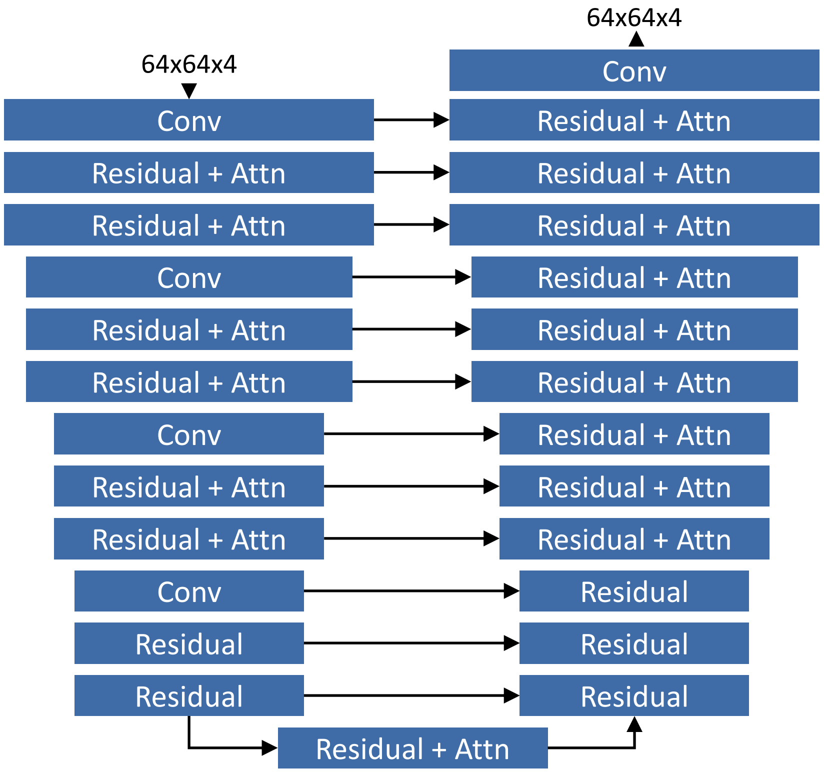
Stable Diffusion 1.5: U-Net Model

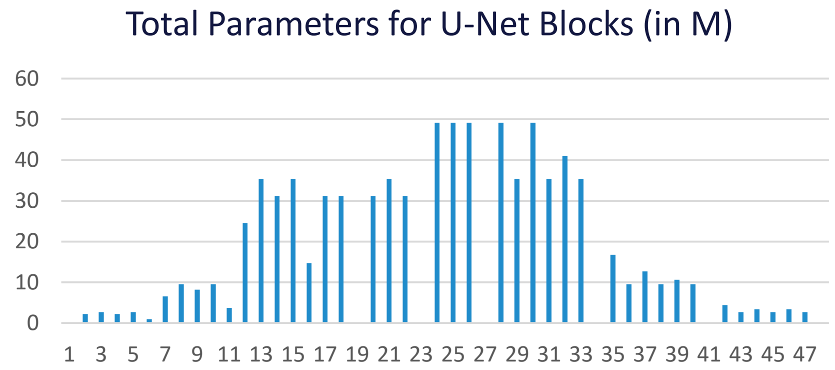
- Latent Diffusion Models (e.g., Stable Diffusion 1.5) are built around a transformer-based U-Net core
- U-Net in Stable Diffusion uses a text-conditioned latent to (re)generate from noise an image with the salient features encoded by the latent
- SD 1.5's U-Net entails 865M parameters and 750B operations

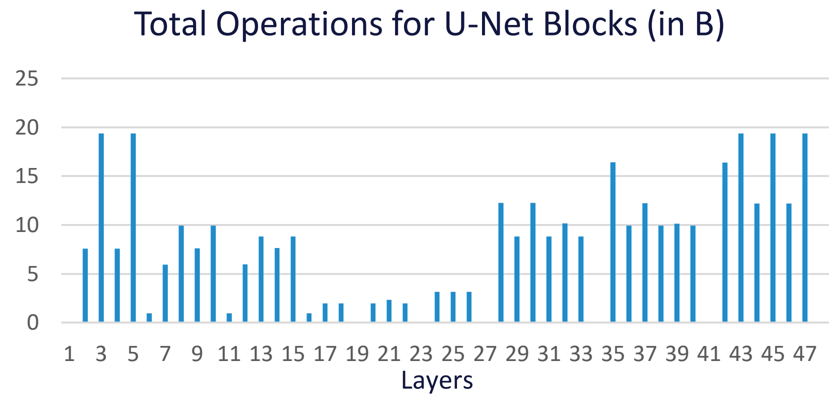


Chambon et al 2022

Stable Diffusion U-Net: Compute vs Parameter Distribution

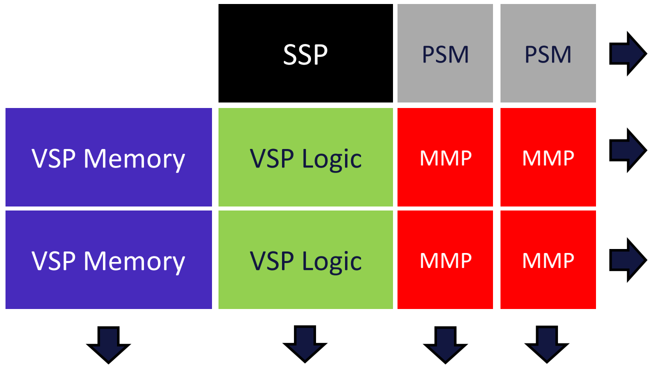






About Expedera

- Packet-based Origin NPU IP focused on edge inference
- Market-validated and production-proven
  - 10M+ devices shipped with Expedera IP
  - Multiple consumer device, smartphone, and automotive production licensees
- Market-leading performance, power, area & latency
  - Support for visual, audio, and generative AI models
  - Single core scales from 3 GOPS to 128 TOPS
  - Customized to use cases



Origin Architecture: 5 fundamental building blocks

Conclusions

- The versatility of LLMs in handling and coordinating different types of data makes them an effective way of processing vision
  - Nearly all image generators are already built on LLM architecture
  - Embodied agents incorporating LLM architecture demonstrate improved reasoning with visual inputs
- The path ahead for LLMs in vision is likely not uniformly transformer-based
  - Transformers lead in image generation; non-transformer models are leading for embodied agents and are less resource-intensive
- Dedicated “brand” or manufacturer support, especially at the hardware level, is necessary to move the capabilities of these models to the edge productively

Resources

Summit & Alliance Resources

- Visit us at booth #322
- Alliance website
  - https://www.edge-ai-vision.com/companies/expedera/

Expedera Resources

- Company Website
  - http://www.expedera.com/
  - White papers, technical briefs, webinars, other
- Pre-silicon PPA Estimations
  - Want cycle-accurate PPA numbers for your use case(s) well before silicon?
  - info@expedera.com
- Contact us directly
  - info@expedera.com

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Anil, R., Borgeaud, S., Wu, Y., Alayrac, J. B., Yu, J., ... & Ahn, J. (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., ... & Ramesh, A. (2023). Improving image generation with better captions. *https://cdn.openai.com/papers/dall-e-3*

Chambon, P., Bluethgen, C., Langlotz, C. P., & Chaudhari, A. (2022). Adapting pretrained vision-language foundational models to medical imaging domains. *arXiv preprint arXiv:2210.04133*.

Dao, T. (2023). FlashAttention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

de Zarzà, I., de Curtò, J., Roig, G., & Calafate, C. T. (2023). LLM multimodal traffic accident forecasting. *Sensors*, *23*(22), 9225.

Fu, D. Y., Dao, T., Saab, K. K., Thomas, A. W., Rudra, A., & Ré, C. (2022). Hungry hungry hippos: Towards language modeling with state space models. *arXiv preprint arXiv:2212.14052.*

Lin, J., Du, Y., Watkins, O., Hafner, D., Abbeel, P., Klein, D., & Dragan, A. (2023). Learning to model the world with language. *arXiv preprint arXiv:2308.01399.*

McKinsey & Co. (2024). *GenAI — The Next S-Curve for the Semiconductor Field. Future of Compute Webinar Series.*

OpenAI. (2023). GPT-4V(ision) System Card. *https://cdn.openai.com/papers/GPTV_System_Card.pdf*

Pirchai, S. & Hassabis, D. (2024) Our next-generation model: Gemini 1.5. *https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/*

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., ... & Rombach, R. (2023). SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PMLR.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10684-10695).

RunwayML. (2022). Stable Diffusion 1.5. *https://huggingface.co/runwayml/stable-diffusion-v1-5*

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., ... & Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, *35*, 36479-36494.

Singh, S., Dewangan, S., Krishna, G. S., Tyagi, V., Reddy, S., & Medi, P. R. (2022). Video vision transformers for violence detection. *arXiv preprint arXiv:2209.03561*.

Srivastava, H., Bharti, A. K., & Singh, A. (2023). Context-Aware Vision Transformer (CaViT) for Satellite Image Classification. *Available at SSRN 4673127*.

Van, M. H., Verma, P., & Wu, X. (2024). On Large Visual Language Models for Medical Imaging Analysis: An Empirical Study. *arXiv preprint arXiv:2402.14162*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wayve (2024). LINGO-2: Driving with Natural Language. https://wayve.ai/thinking/lingo-2-driving-with-language/