HAILO

Advanced Analytics for VMS Platforms

Avi Baum & Mark Grobman

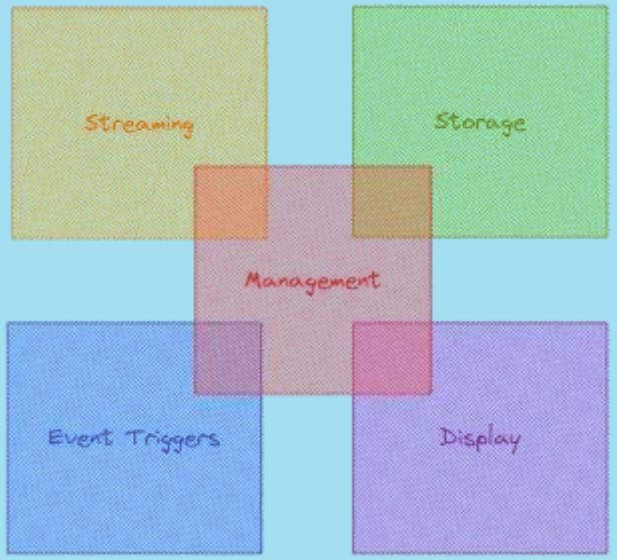
7 February 2023

Agenda

- VMS key components
- Analytics in VMSs
- Deployment topologies
- Benefit of analytics (by example)
- The rise of transformers

VMS in a nutshell

- Video Management System (VMS) as a hypernym for all functional components involved in multiple cameras systems
- In this talk context, VMS encapsulates all the **key components** of this system
- Mostly deployed in the context of security systems
- Applicable in any massive deployment of video sources

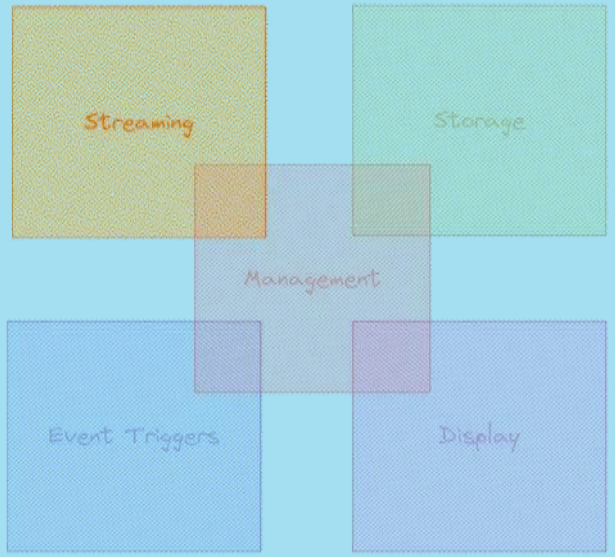


VMS Key Components: Streaming

- Determines what content is being transferred and when

- Key factors
  - → No. of streams / channels
  - → Per channel frame rate
  - → Encoding capacity
  - → Latency
  - → Bandwidth

- Typical configurations
  - → Streaming all
  - → Store & forward
  - → Stream upon event



$$Actual\ Bandwidth \sim N \cdot \sigma(W \cdot H \cdot FPS) \leq Network\ BW$$

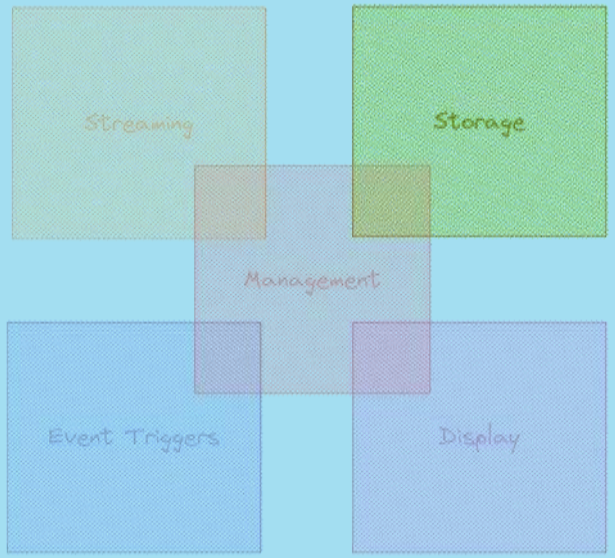
$$N \cdot W \cdot H \cdot FPS \leq Encoder\ BW$$

VMS Key Components: Storage

- Determines what content is being stored and when

- Key factors
  - → No. of streams / channels
  - → Per channel frame rate
  - → Storage duration
  - → Storage space
  - → Read & Write speed
  - → Efficient indexing & search

- Typical configurations
  - → Local storage
  - → Central storage



$$Actual\ Bandwidth \sim N \cdot \sigma(W \cdot H \cdot FPS) \leq Write\ Speed$$

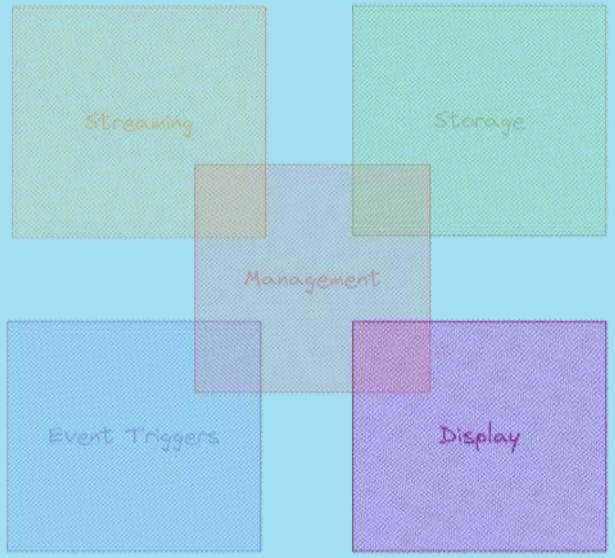
$$Duration \sim N \cdot \frac{\sigma(W \cdot H \cdot FPS)}{Storage\ space}$$

VMS Key Components: Display

- Determines what content is being displayed and when

- Key factors
  - → Number of displays
  - → Minimal stream resolution
  - → Display resolution
  - → Decoding speed
  - → Operator attention (which streams are selected)

- Typical configurations
  - → Co-located with management entity
  - → Co-located with storage entity
  - → Separate



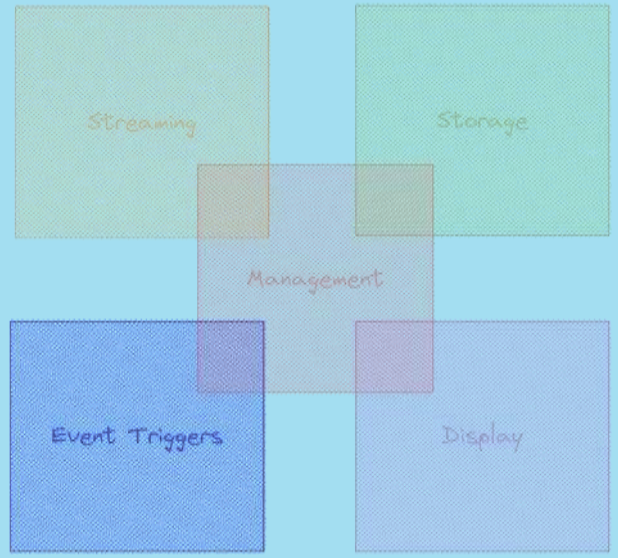
$$\frac{D \cdot W \cdot H}{Stream_W \cdot Stream_H} \cdot Refresh\_Rate \leq Decoding\ speed$$

VMS Key Components: Events

- Triggers events from selected video sources

- Key factors
  - → No. of streams/channels
  - → Per channel frame rate
  - → Decoding capacity
  - → Event rate
  - → Latency
  - → Accuracy (low miss-rate ; low false alarms)

- Typical configurations
  - → Camera-attached
  - → Gateway-bound



$$Actual\ TOPS \sim N \cdot \frac{OPS}{frame} \cdot FPS \leq Available\ TOPS$$

$$N = min\left(\frac{Decoder\ BW}{W \cdot H \cdot FPS}, \frac{OPS \cdot FPS}{frame}\right)$$

Where to apply analytics?

- In what stages of the VMS processing pipeline? .. All
- Do they serve the same purpose? .. No

- Motivation
  - → Camera-attached analytics ➔ Lower per-channel bandwidth
  - → Gateway bound analytics ➔ Improve latency by offloading central processing
  - → Storage entity analytics ➔ Limit storage bandwidth and capacity
  - → Display entity analytics ➔ Display relevant activity
  - → Event triggering ➔ Lower / better balance load on other entities

Why is advanced analytics needed?

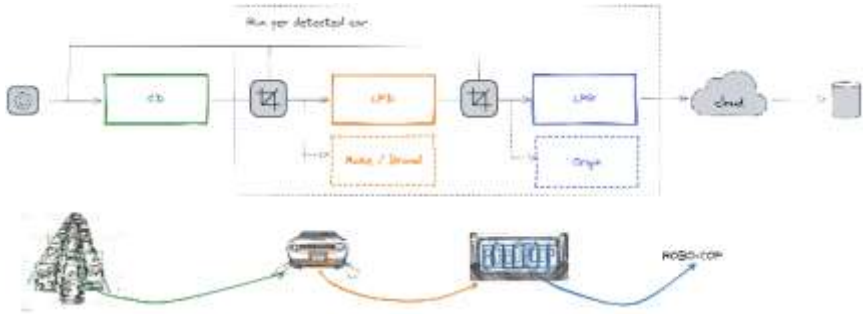
- Why do we need to make a distinction between “basic” and “advanced”?

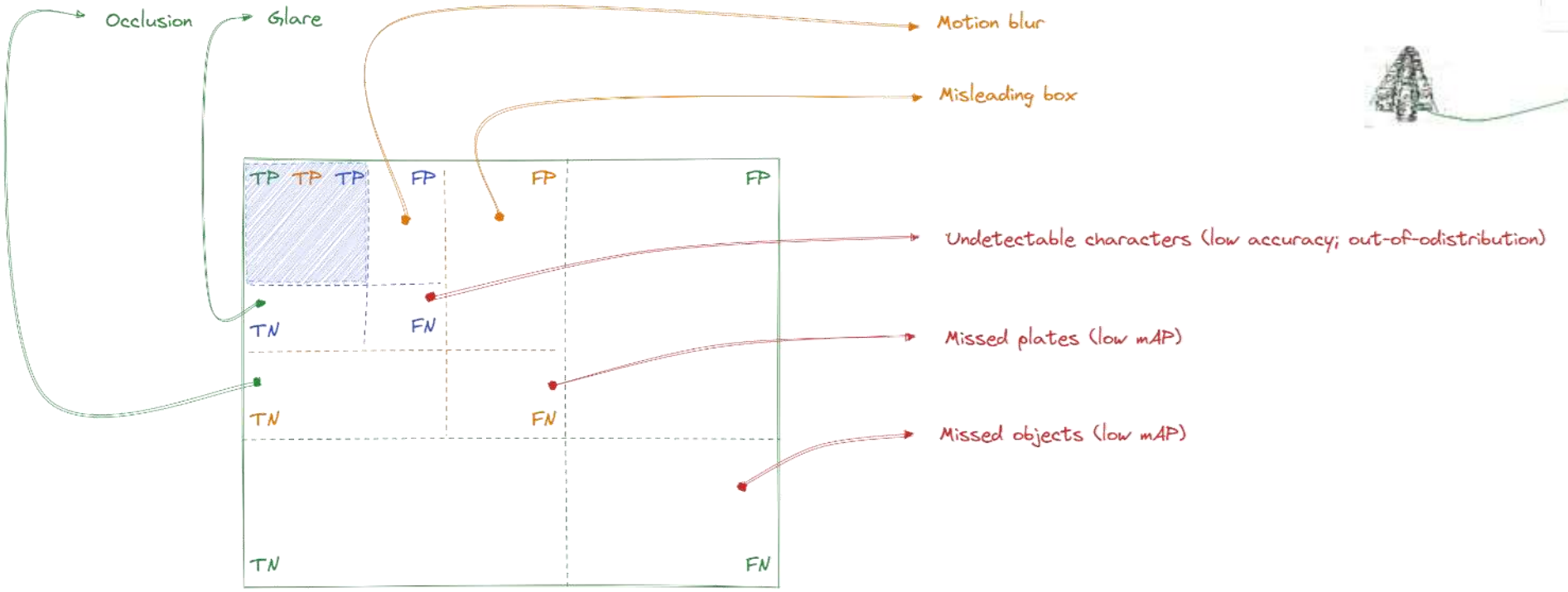
- What is **advanced** analytics?

- Several options to define
  - → More .. Functionality — running more analytic functions in parallel
  - → Better.. Performance — higher true positives with lower false alarms and miss rates
  - → Higher.. Density — more channel per system

Analytics was used to be a bottleneck, this is no longer the case

What does better analytics enable? ALPR Example





CD = Car Detector

LPD = License Plate Detector

LPR = License Plate Recognizer

CD:
- more classes => improve FN
- allow car make/brand (feed-forward)
- car embeddings

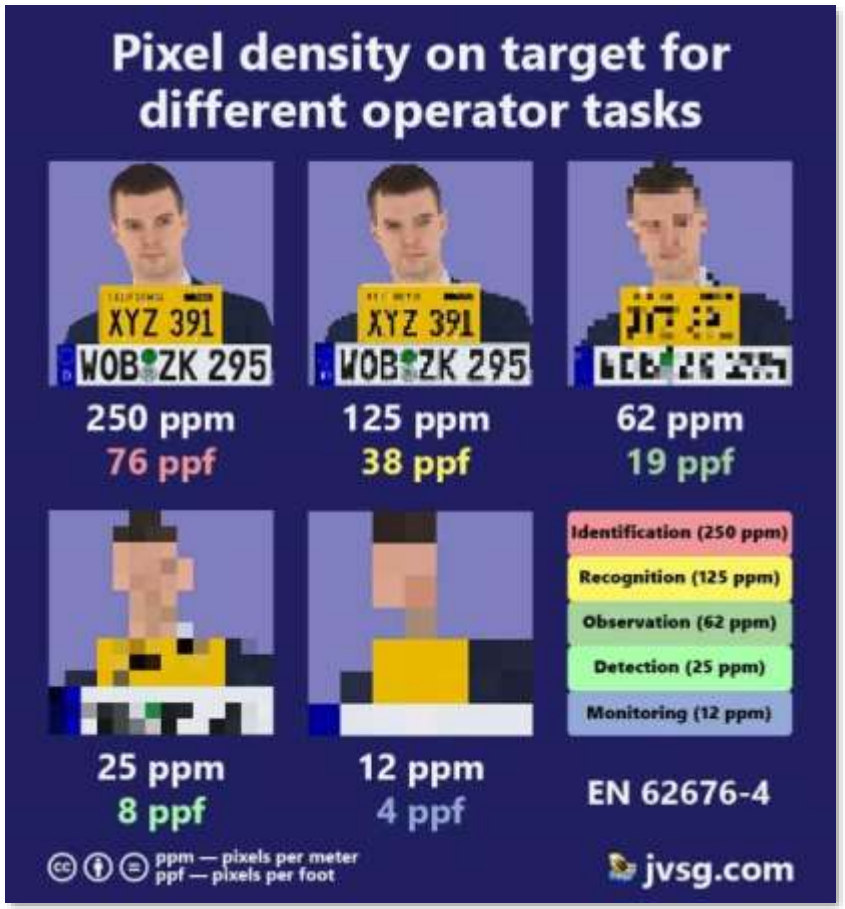
LPD:
- more variety of plates => improve FP
- allow car origin (feed-forward)

LPR:
- locale adaptation
- beyond character recognition
- image artifacts due to environment (blur; glare/reflections...)

Advanced analytics requires more **processing capacity** to provide **better performance**

Standardization of quality

- Domain-specific standards for image quality requirements
  - → IEC EN62676 (Image Quality for Video Surveillance Systems)
- Establishes a baseline for required analytics
  - → Representing nominal conditions only (not 'in the wild')
  - → Lacking the insights of common AI perception test paradigms (e.g. a well curated dataset)
- Advanced analytics can be used to
  - → **Meet** standard defined functionality in **extreme** conditions
  - → **Meet** standard functionality with **simpler** endpoint
  - → **Improve** over standard baseline at **same** conditions



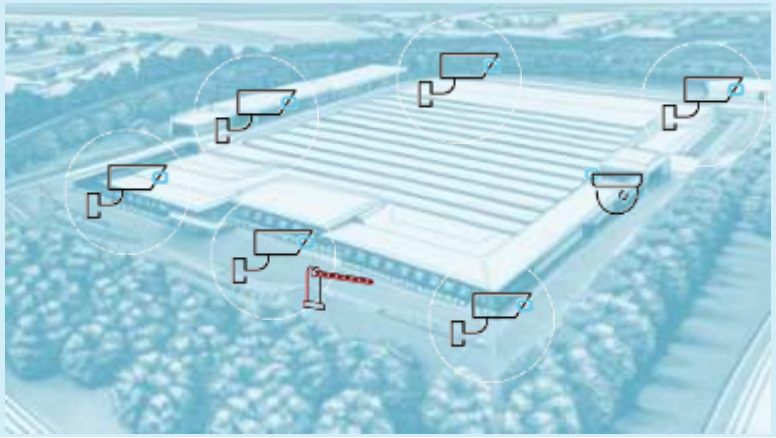
Advanced analytics to **extend** operational envelope

Analytics in Practice

Typical System Configurations

Multiple Topologies

Campus Perimeter



Campus Indoor

Server Room

Security Room

Analytics in context

Few typical configurations for introducing analytics to a VMS system

1. Standalone analytics box / edge box

2. Storage-attached analytics

3. Server-attached analytics

Video Analytics Box

- Purpose built box
- Receives encoded streams
- Handles per-stream analytics to lower load on next stages
- Price pressured
- No local display
- No local storage

- Main KPIs
  - → Decoding capacity @ MPPS
  - → TOPS / frame
  - → Price: $ / channel

- Measured data on RSC-101:
  - CPU Intel Celeron J6413
  - AI Hailo-8
- Performance

Decoding (2MP)	570 fps
TOPS / channel	~ 0.75
➔ Total channels	16
Price	25 $/channel

Storage Server

- High capacity storage
- Workstation class
- Decoding limited platform

- Typical pipelines
  - → Background indexing
  - → Minimal real-time triggers

- Main KPIs
  - → Recording capacity
  - → Video Metadata (VMD) Indexing capacity
  - → Price: $ / channel

- Example Based on Premio FlacheSAN1N36M
  - CPU AMD Epyc
  - NVMe / AI x36 M.2
- Performance

Traffic (10 Gb)	< 300 channels
Recording	48 hr / channel
Analytics	Configurable 1 : 1 → 1 : GOP
→ Total channels	256

Management Server

- Server class host
- Heterogeneous configuration (CPU + GPU + NPU)

- Typical pipelines
  - → Video metadata
  - → Real-time event triggers
  - → Visualization overlay
- Main KPIs
  - → Display capacity
  - → Triggering latency

- Lenovo SR630-V2
  - CPU — Xeon Gold
  - AI — 2x Falcon H-8 (200-300 TOPS)
- Performance

Decoding (2MP)	1600 fps
Display (4K @30fps)	4
Analytics (TOPS/ch)	1.5
➔ Total channels	64 @ 25 / 48 @ 30

Case Study I: Accuracy / Minimize false triggering

- Accuracy is not just an academic KPI
- It translates into concrete VMS value in the event triggering entity

  → Lower disk space ⇔ more channels for same platform

- KPI:

  → **x4** on capacity (16 → 64 channels)

  → **85%** <u>reduction</u> on false alarms

Case Study 2: Load balancing / Offloading

- **Offloading** main processing entity
  - → Forwarding only relevant data crop
- Lowering **latency** to enable shorter RTT
- Better **privacy**,
  - → Forwarding only cropped data
  - → Information distribution
  - → Storage duration restrictions (GDPR)

Case Study 3: Scalability

- A well-balanced system gives the opportunity for consistency across scales
- AI analytics is no longer a bottleneck

Trending up – better analytics, for more reasons

- Trend #1: Mixing text & video
  - → Result of the recent evolution in **transformers**
  - → Enabling effective mixture of vision & NLP
- Trend #2: More Analytics
  - → Leverage analytics beyond event triggering
    - Lower streaming bandwidth
    - Improve storage indexing
    - Enhance display capabilities
    - Equip management entity with perception advantages
  - → Enabled by the lower TCO of advanced analytics (less $ / function / channel)

The Rise of Transformers

What Are Transformers?

- Introduced in 2017 in “Attention Is All You Need / Vaswani et. al.”

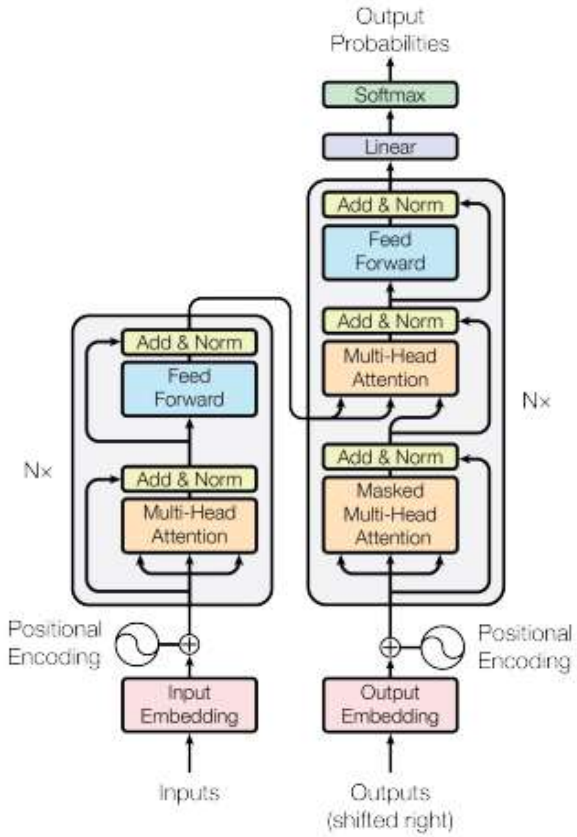
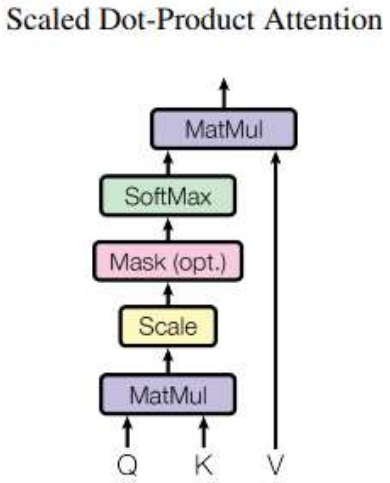


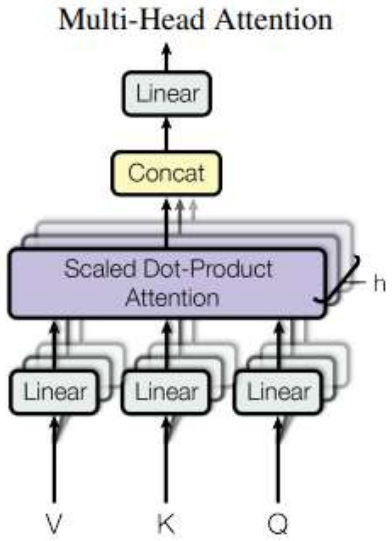
Figure 1: The Transformer - model architecture.

What Are Transformers?

- Introduced in 2017 in "Attention Is All You Need / Vaswani et. al."
- New building block – Multi-Head Attention (MHA)

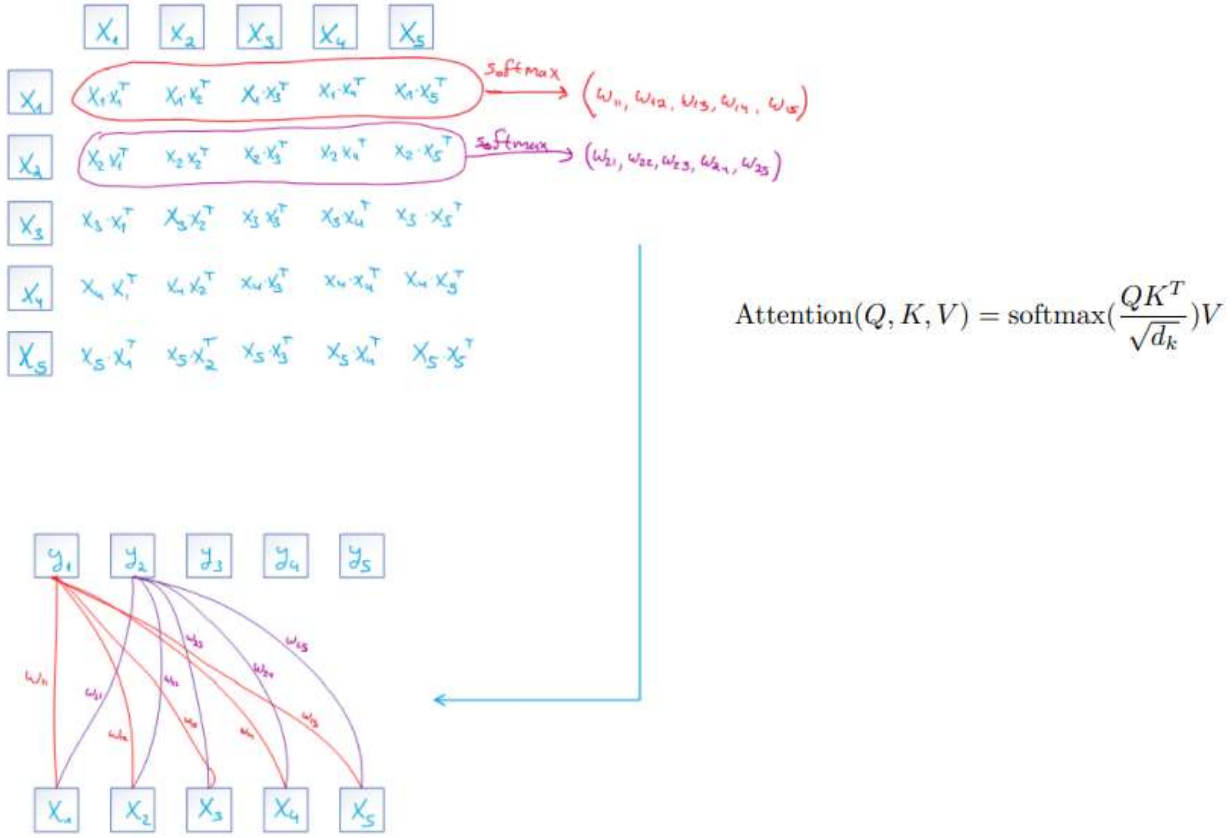
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$





What Are Transformers?

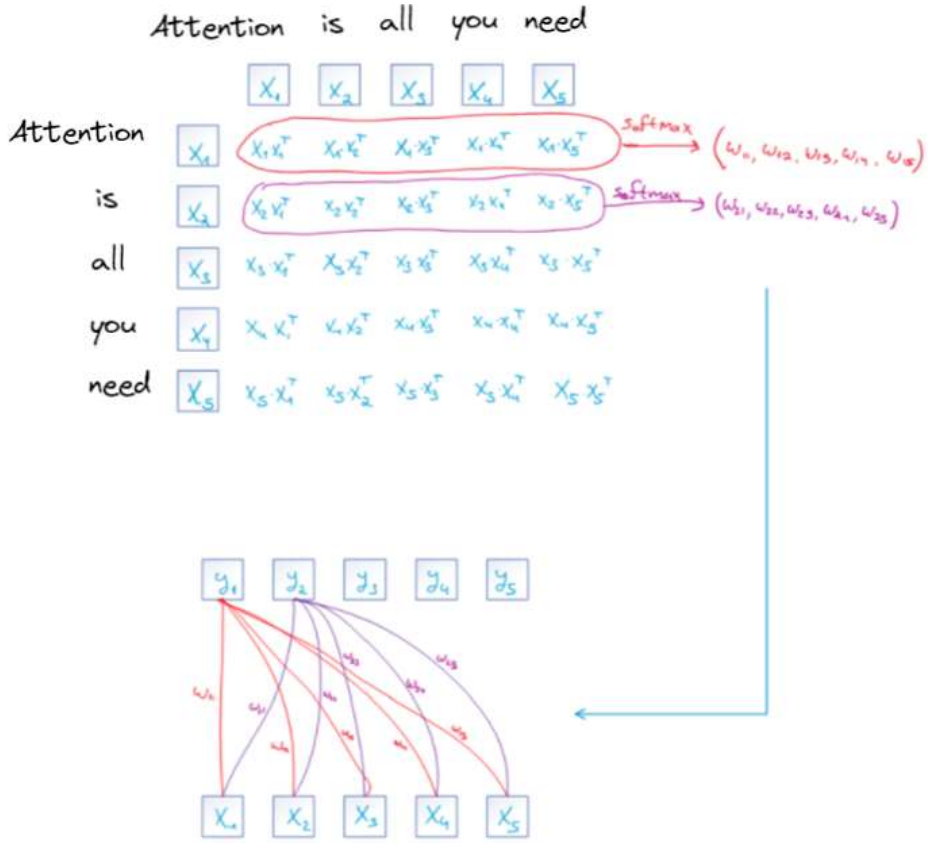
- Introduced in 2017 in "Attention Is All You Need / Vaswani et. al."
- New building block – Multi-Head Attention (MHA)
- MHA is:
  - → Global (Context-aware)
  - → Dynamic (Data-driven)



$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

What Are Transformers?

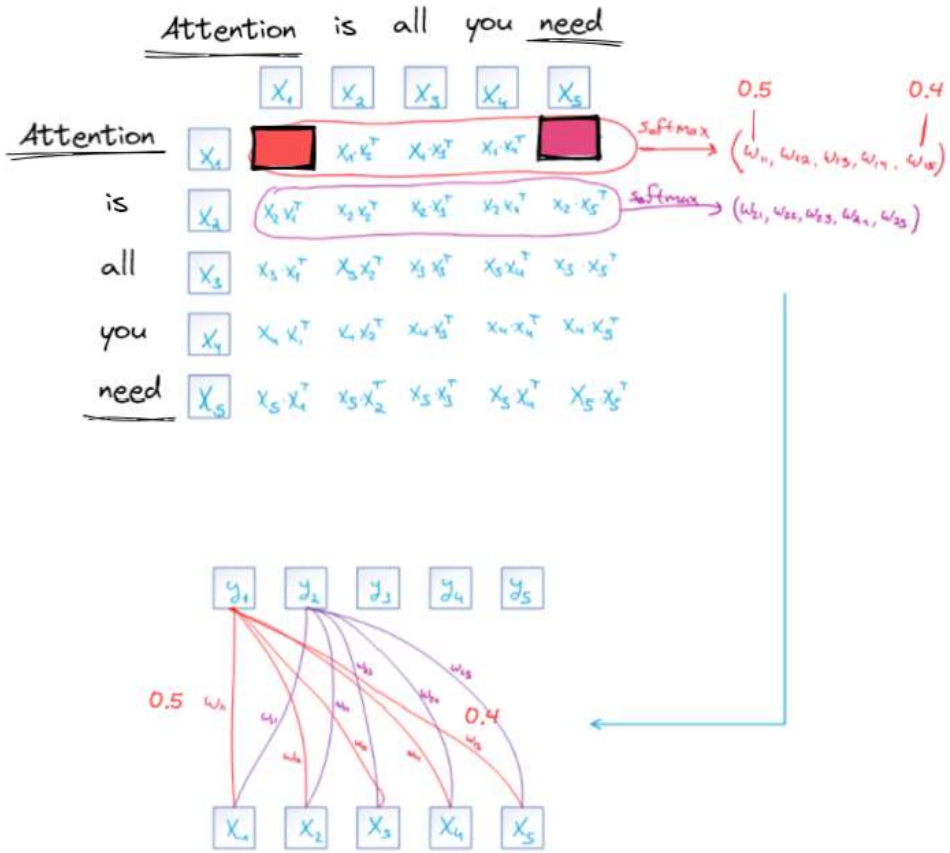
- MHA is:
  - → Global (Context-Aware)
  - → Dynamic (Data-driven)



$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

What Are Transformers?

- MHA is:
  - → Global (Context-Aware)
  - → Dynamic (Data-driven)

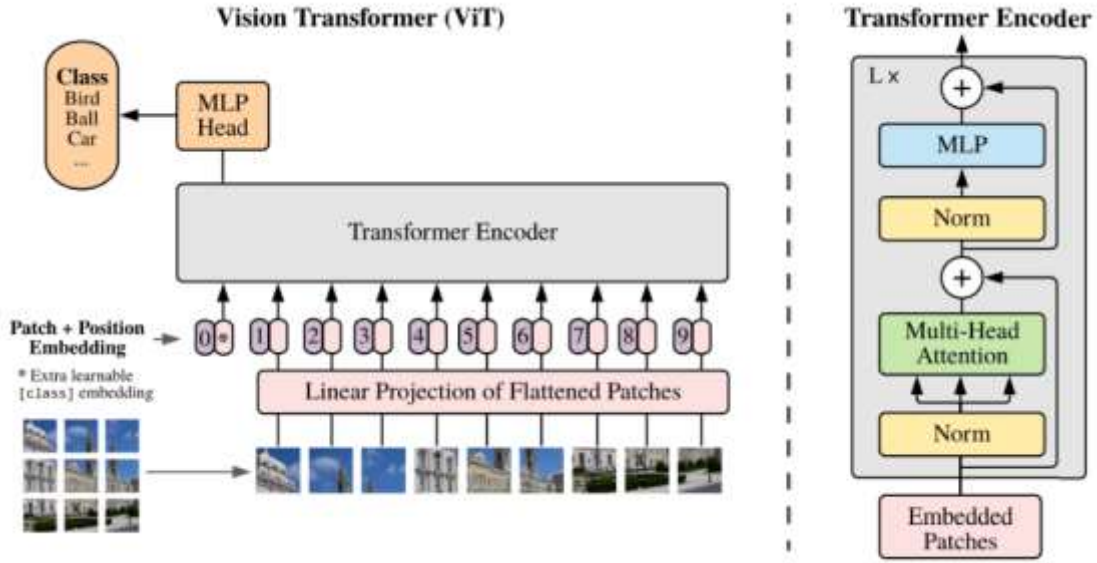


$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Transformers for computer vision

- Introduced in 2017 in "Attention Is All You Need / Vaswani et. al."
- Introduced for NLP but generalized across tasks – vision in particular

  → An Image Is Worth 16X16 Words / Dosovitskiy et al. 2020



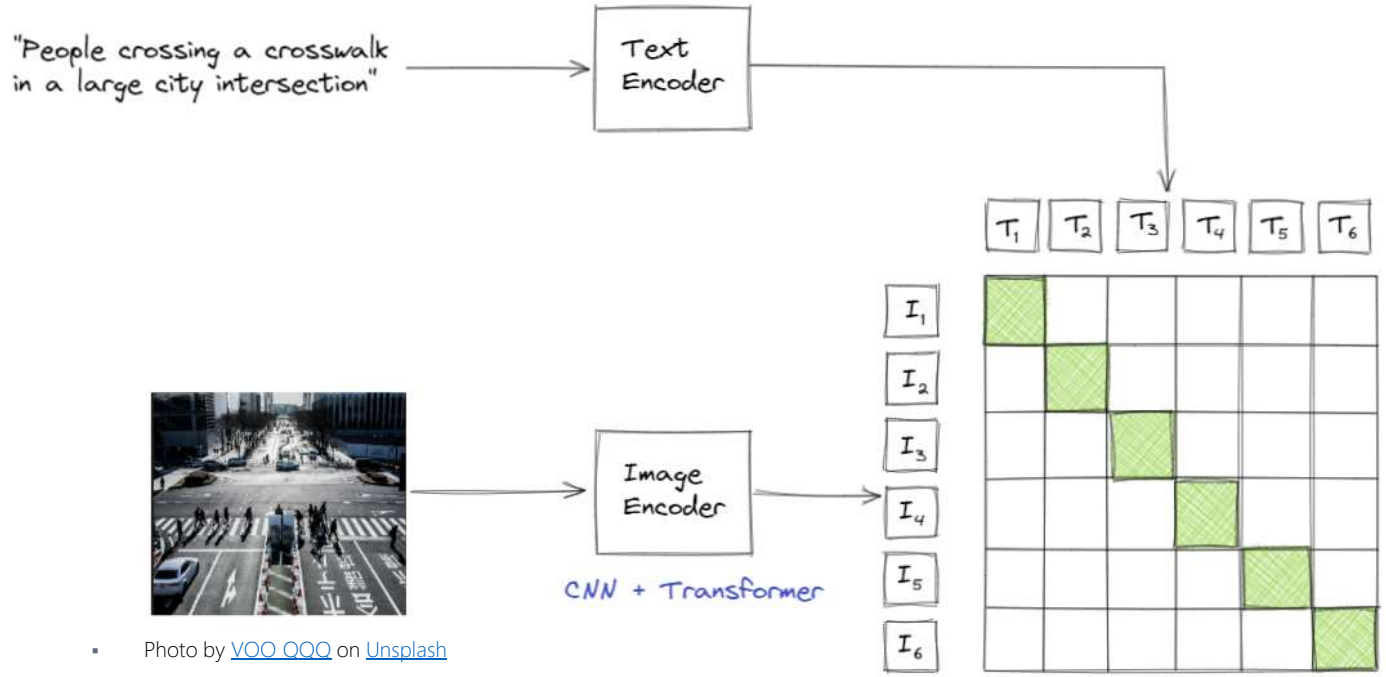
Why Transformers?

- Why are Transformers so popular?
  - → Scale well with data
  - → Highly parallel and efficient to train
  - → Simple architectures - continue the trend of less domain expertise
  - → Make it easy to **fuse across domain/modalities**
  - → SoTA

Why Transformers?

- Why are Transformers so popular?
  - → Scale well with data
  - → Highly parallel and efficient to train
  - → Simple architectures - continue the trend of less domain expertise
  - → Make it easy to **fuse across domain/modalities**
  - → SoTA
- Are CNN's dead?
  - → No! Strong priors on the task translate to **efficiency**.
  - → Still the choice for small/mid range models
- Best of both worlds: CNN +Transformer

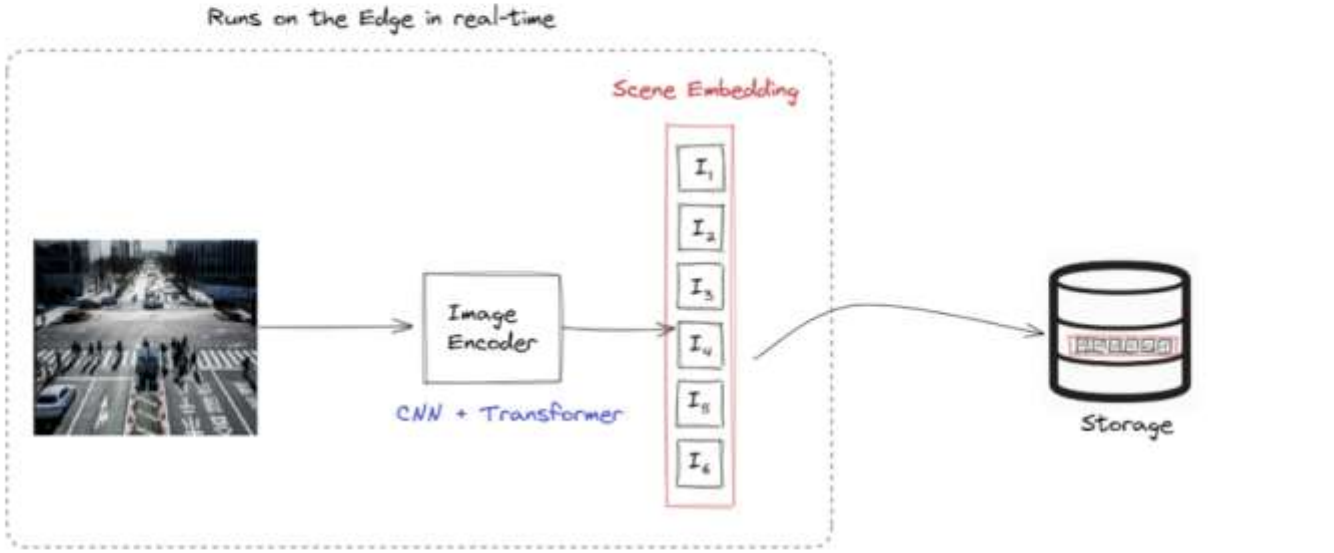
Language + Vision unlocks semantic search



- Photo by VOO QQQ on Unsplash

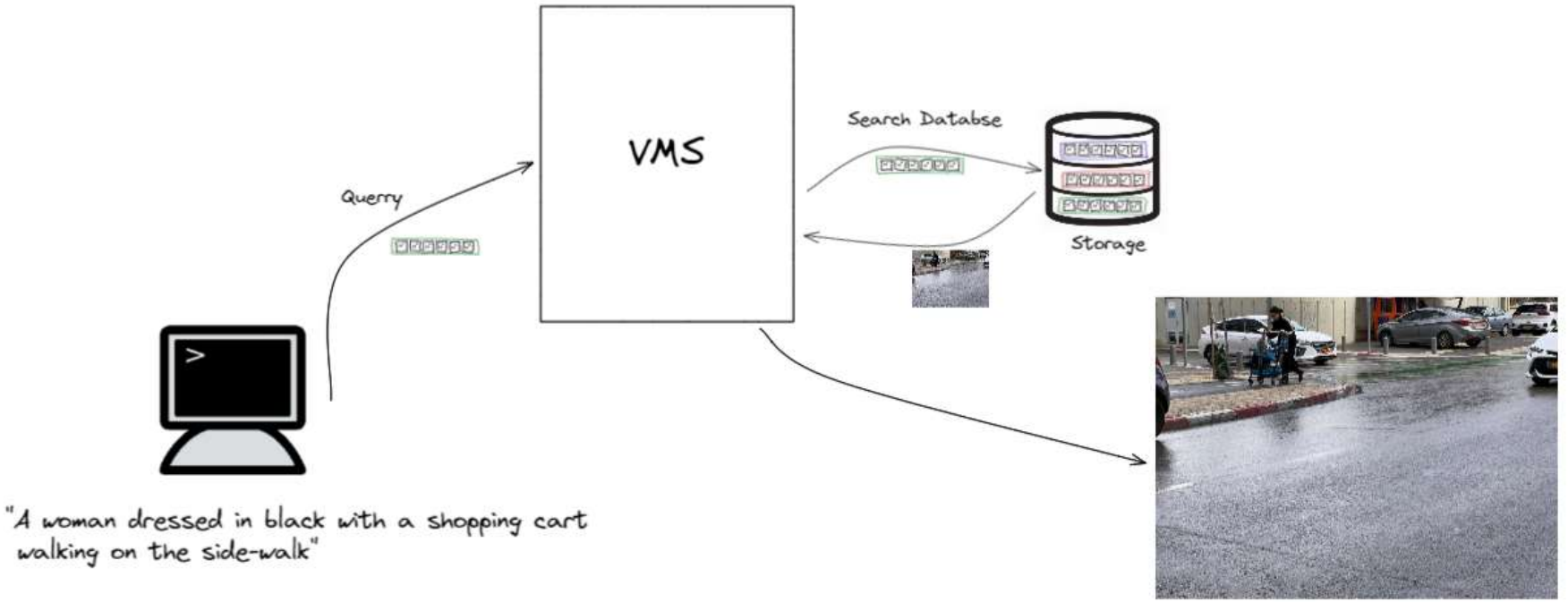
Training Stage - Train on images with captions using contrastive learning.

Language + Vision unlocks semantic search



Inference Stage - run only the Image Encoder and store the embedding on storage

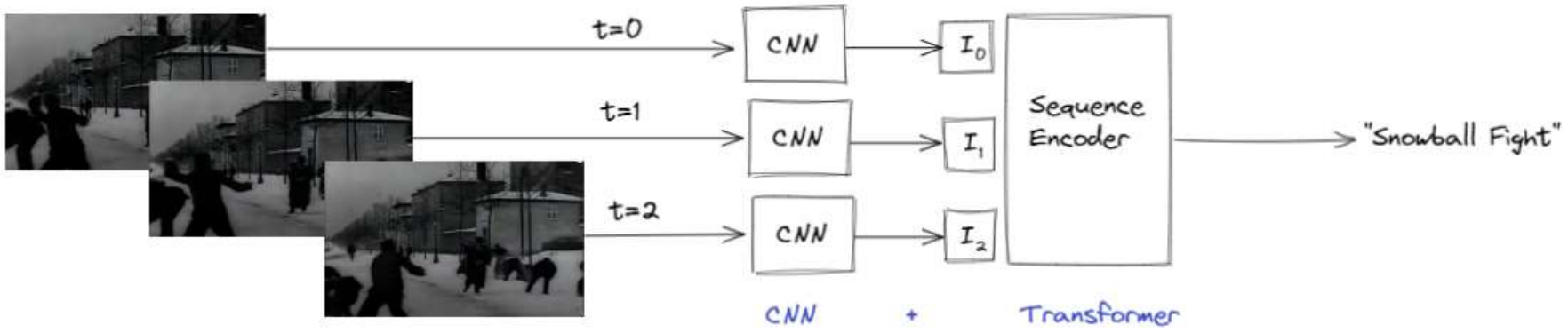
Language + Vision unlocks semantic search



Offline Stage - search in the database using natural language

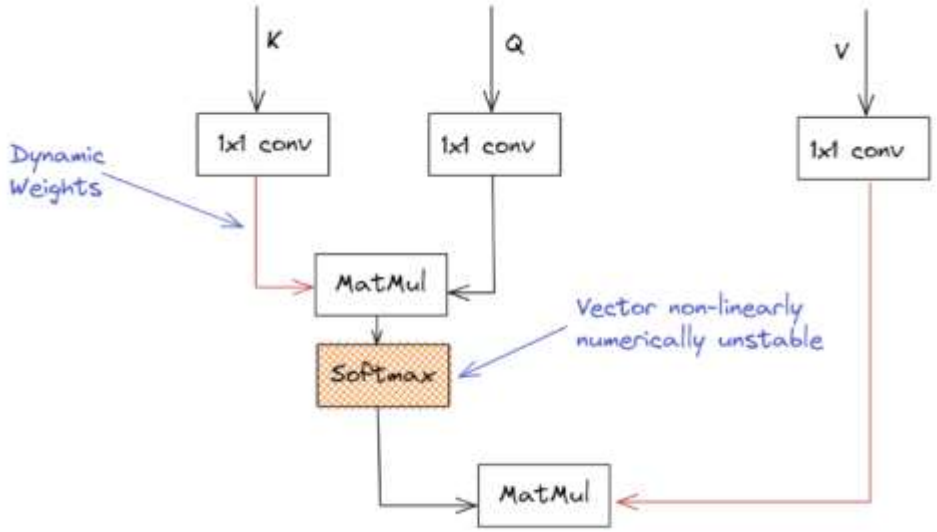
Frame fusion for activity recognition

- Transformers are excellent at working on long sequence
- Can be used to classify activity over long times efficiently.



Transformers acceleration on Hailo

- Transformer acceleration on the edge
  - → Dynamic weights – data driven weight
  - → High Throughput Softmax
    - When using external engine results in bottleneck
    - Implemented directly on NN core
- Transformers have inherent increase latency



Hailo Support For Transformer

- SW Suite 2023-01 includes support for Transformers
  - → MZ release of ViT
  - → First edge AI accelerator able to run ViT-B in real-time.

Model	Embedding	#heads	#layers	Params [M]	Ops [G]
ViT-B	768	12	12	86	17.5

- We believe in transformers on the edge
  - → Intense work over the past two years
  - → Ongoing optimization in all facets of our product

Empowering Product Creators to Harness Edge AI and Vision

The Edge AI and Vision Alliance (www.edge-ai-vision.com) is a partnership of 100+ leading edge AI and vision technology and services suppliers, and solutions providers

Mission: To inspire and empower engineers to design products that perceive and understand.

The Alliance provides low-cost, high-quality technical educational resources for product developers

Register for updates at www.edge-ai-vision.com

The Alliance enables edge AI and vision technology providers to grow their businesses through leads, partnerships, and insights

For membership, email us: membership@edge-ai-vision.com

edge ai + vision ALLIANCE™

Join us at the Embedded Vision Summit
May 22-25, 2023—Santa Clara, California

The only industry event focused on practical techniques and technologies for system and application creators

- *"Awesome! I was very inspired!"*
- *"Fantastic. Learned a lot and met great people."*
- "*Wonderful speakers and informative exhibits!*"

Embedded Vision Summit 2023 highlights:

- **Inspiring keynotes** by leading innovators
- High-quality, practical **technical, business and product talks**
- Exciting **demos**, **tutorials** and **expert bars** of the latest applications and technologies

Visit www.EmbeddedVisionSummit.com to learn more and register

HAILO

Empowering Intelligence

THANK YOU

hailo.ai contact@hailo.ai