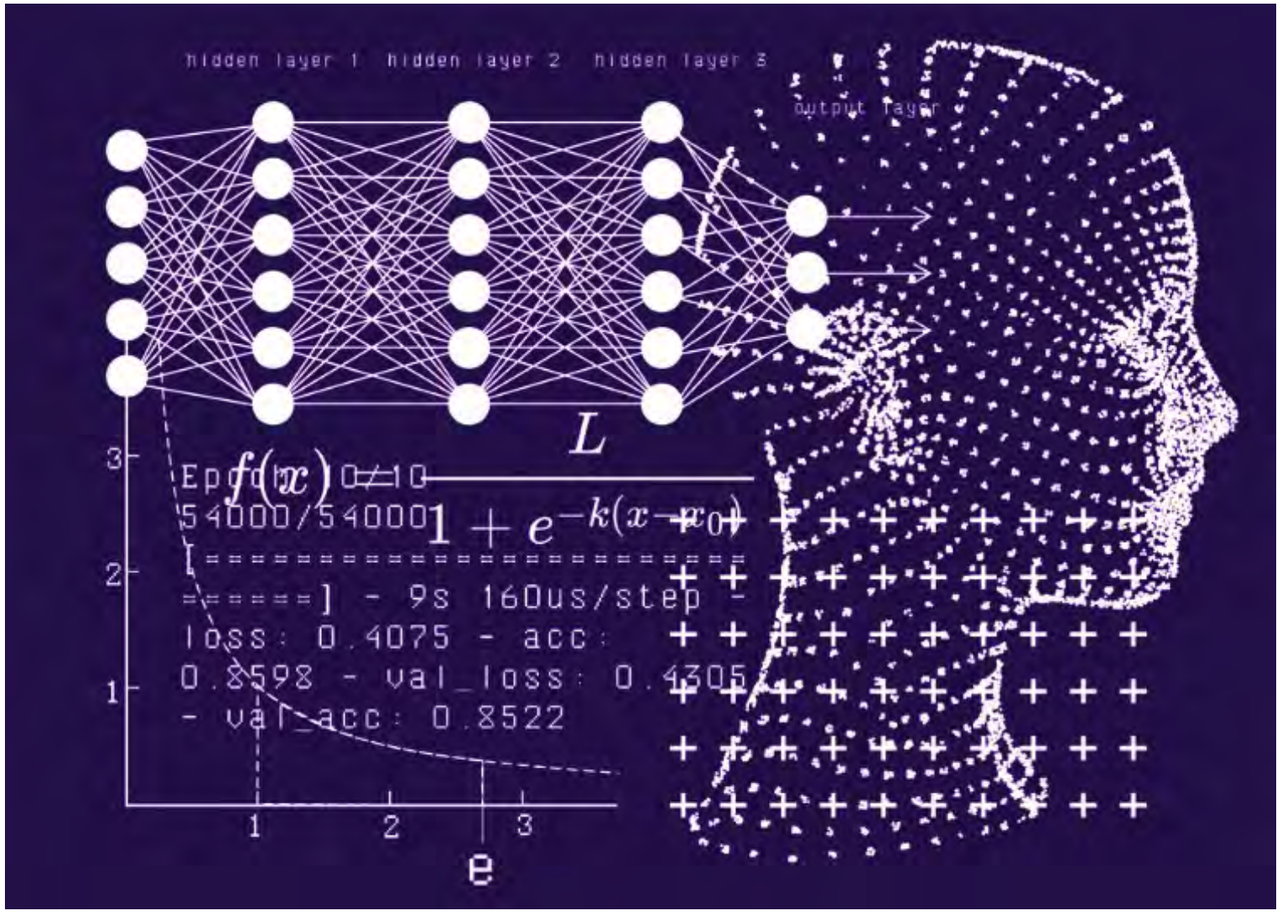
AI-Based Video Codecs and Semantic Video Search

Rob Gonsalves
Engineering Fellow, Avid
February 23, 2022

Using AI/ML for Media Content Creation

- Background of AI and ML
- AI-Based Video Codecs
- Semantic Video Search
- Q&A



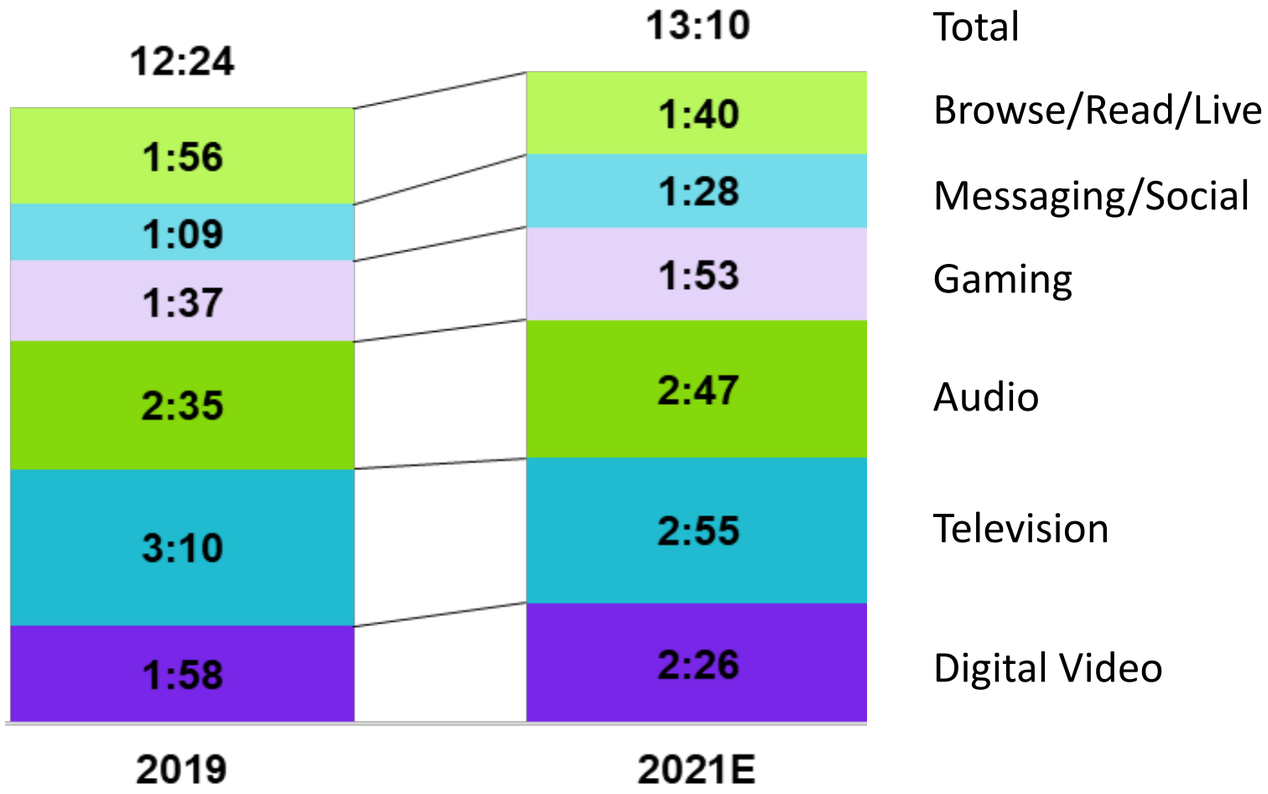
Increased Demand for Media Content

There is a strong, continued consumer demand for content.

Production teams are expected to create more high-quality content using fewer resources.

Video and Audio Combine for 8:08 Daily Attention

Up 25 Min, US adults, Hours:Minutes

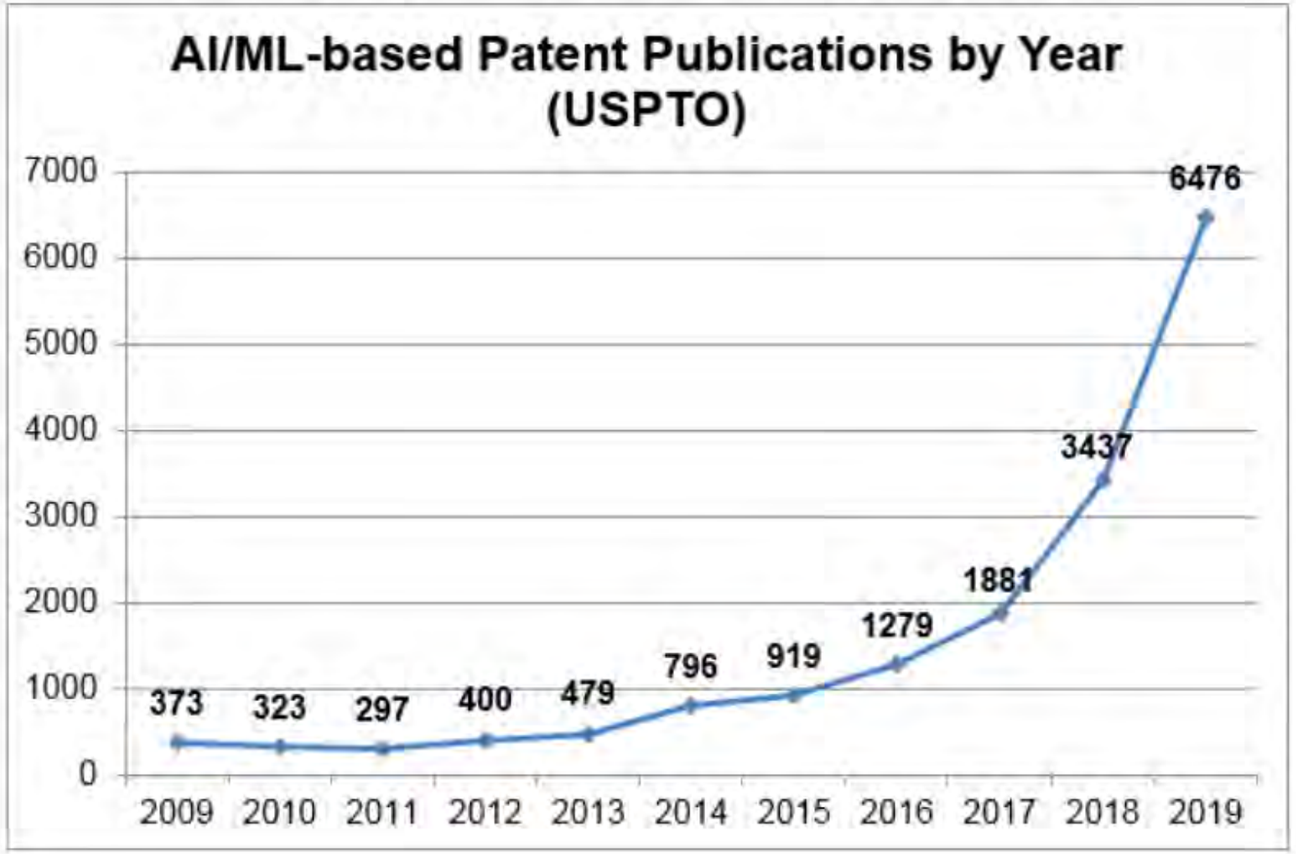


Source: Activate Consulting “Activate Technology and Media Outlook, 2022” (October 2021) Primary sources can be found in that document

Recent Developments in AI/ML

Since 2015, enabling technologies spurred increased research and development of AI/ML

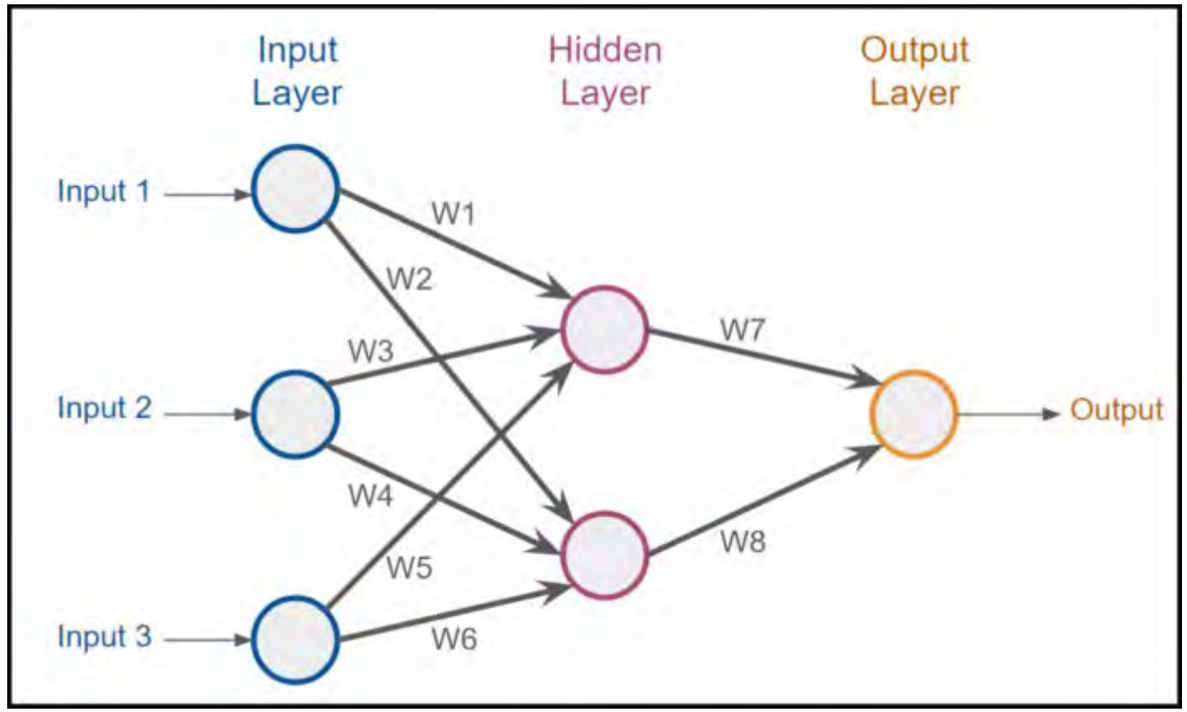
- Development of Deep Learning AI Models and Software Tools
- Availability of Large-scale Datasets
- Use of GPUs for Training and Inferencing



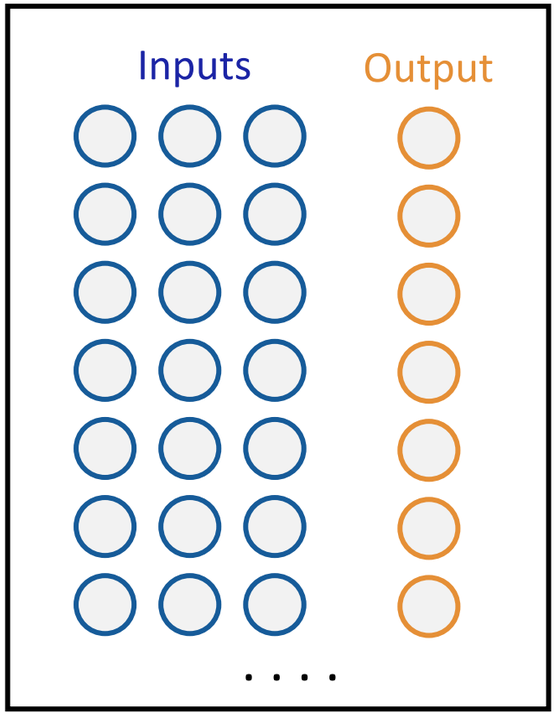
Source: USPTO

Artificial Neural Networks - Basics

Diagram of a Simple ANN



Training Data



Forward Pass (Inferencing)

Backwards Pass (Adjust weights during training)

Using Machine Learning for Media Content Creation

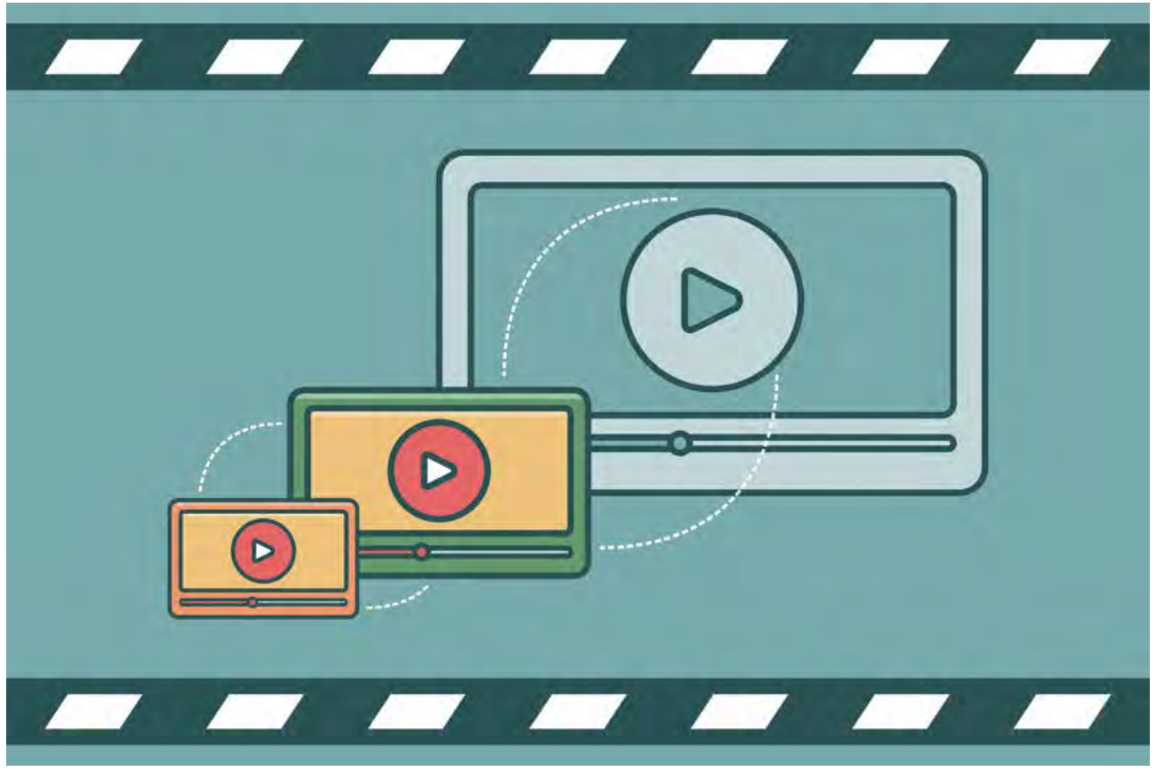
ML Media Applications

- Metadata Extraction
- Speech-to-text
- Content Classification
- Color Correction
- Video super-resolution resize
- Video Compression
- Search & Retrieval
- Etc.

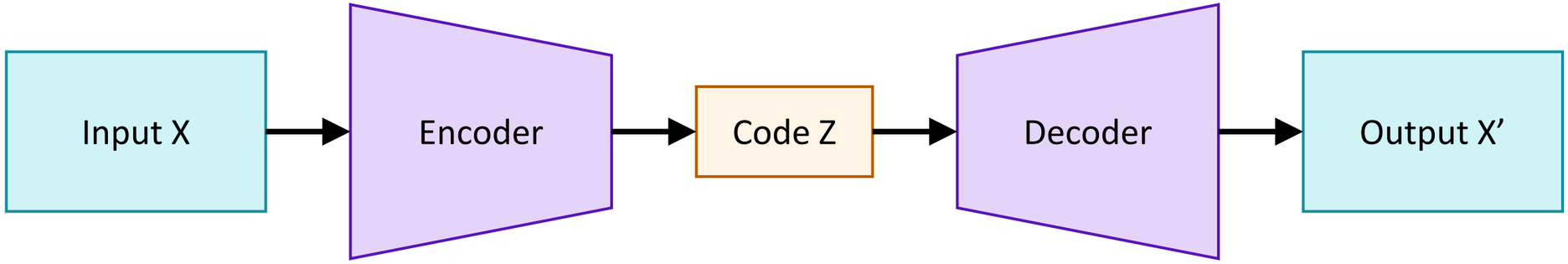
AI-Based Video Codecs

Goals: Improve quality of compressed video and reduce the required bandwidth for production codecs

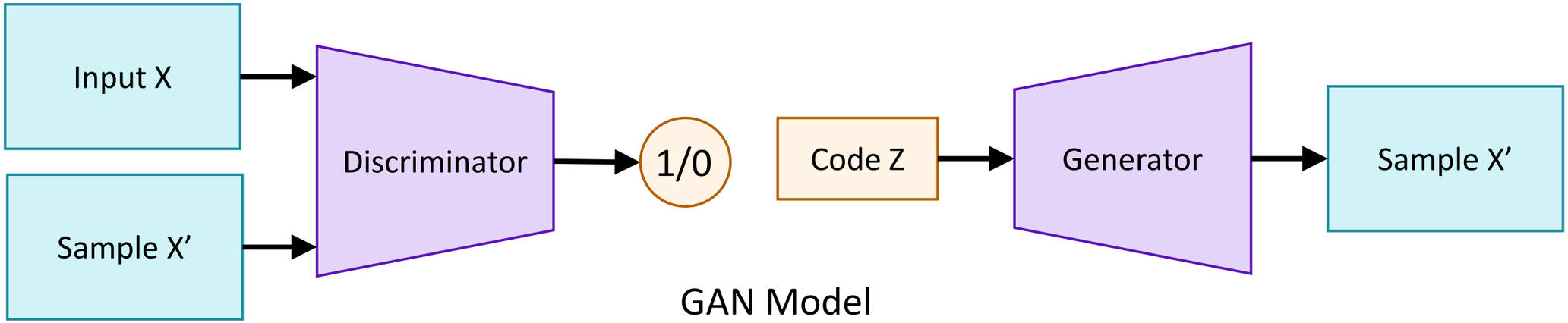
- AI Models for Codecs
- AI-Based Codecs
- Video Quality Metrics
- Results



AI Models for Video Codecs



Auto-Encoder Model



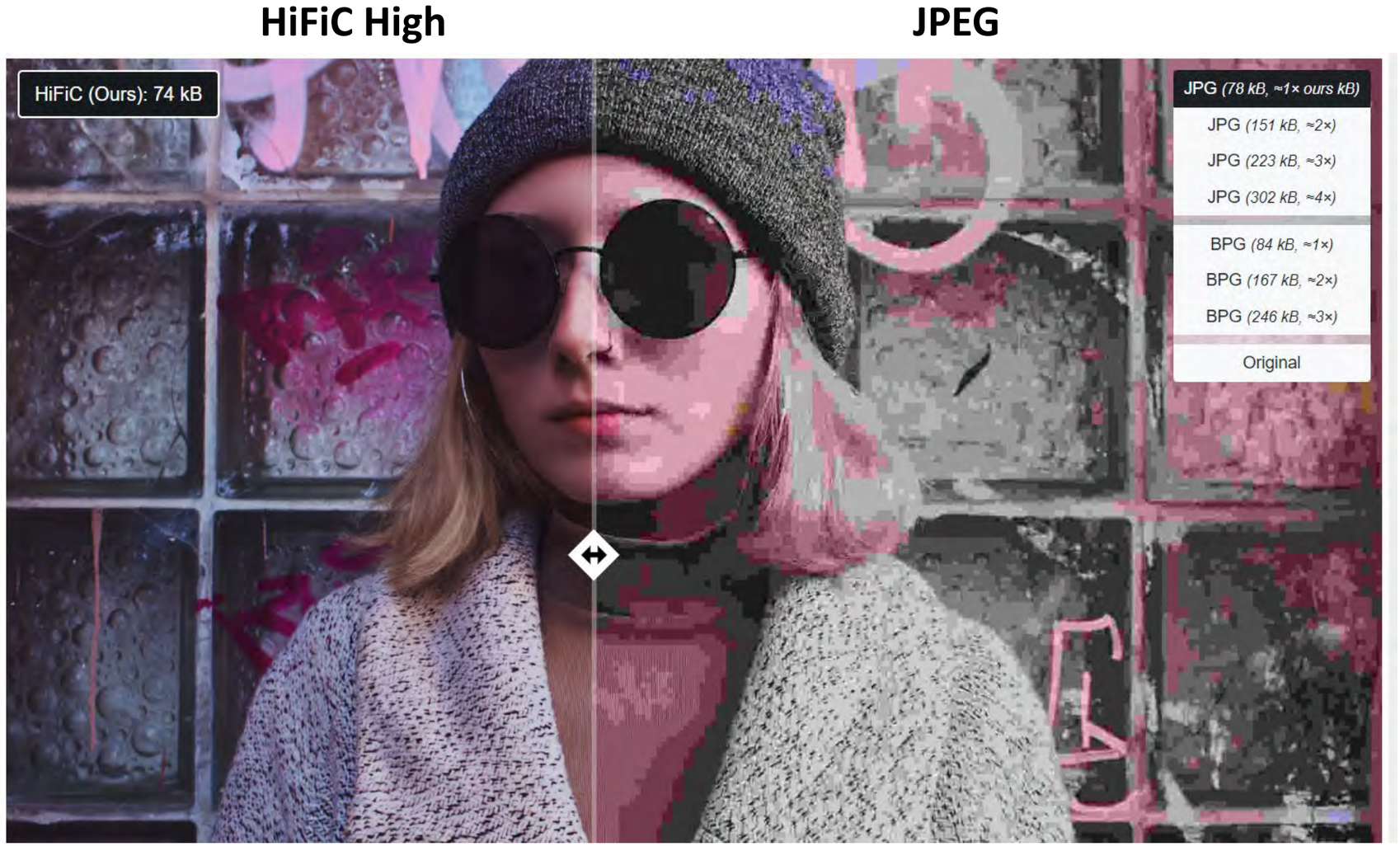
GAN Model

Video Codecs

Class	Codec	Description
End-to-end AI	**HiFiC**	Still image codec from Google
End-to-end AI	**WaveOne**	Learned codec from WaveOne
End-to-end AI	**420 Autoencoder**	Avid research project
AI-Assist	VC-6	Hybrid codec from V-Nova
Procedural	VVC (H.266)	Versatile Video Coding
Procedural	HEVC (H.265)	High Efficiency Video Coding
Procedural	AVC (H.264)	Advanced Video Coding

Google HiFiC Image Codec

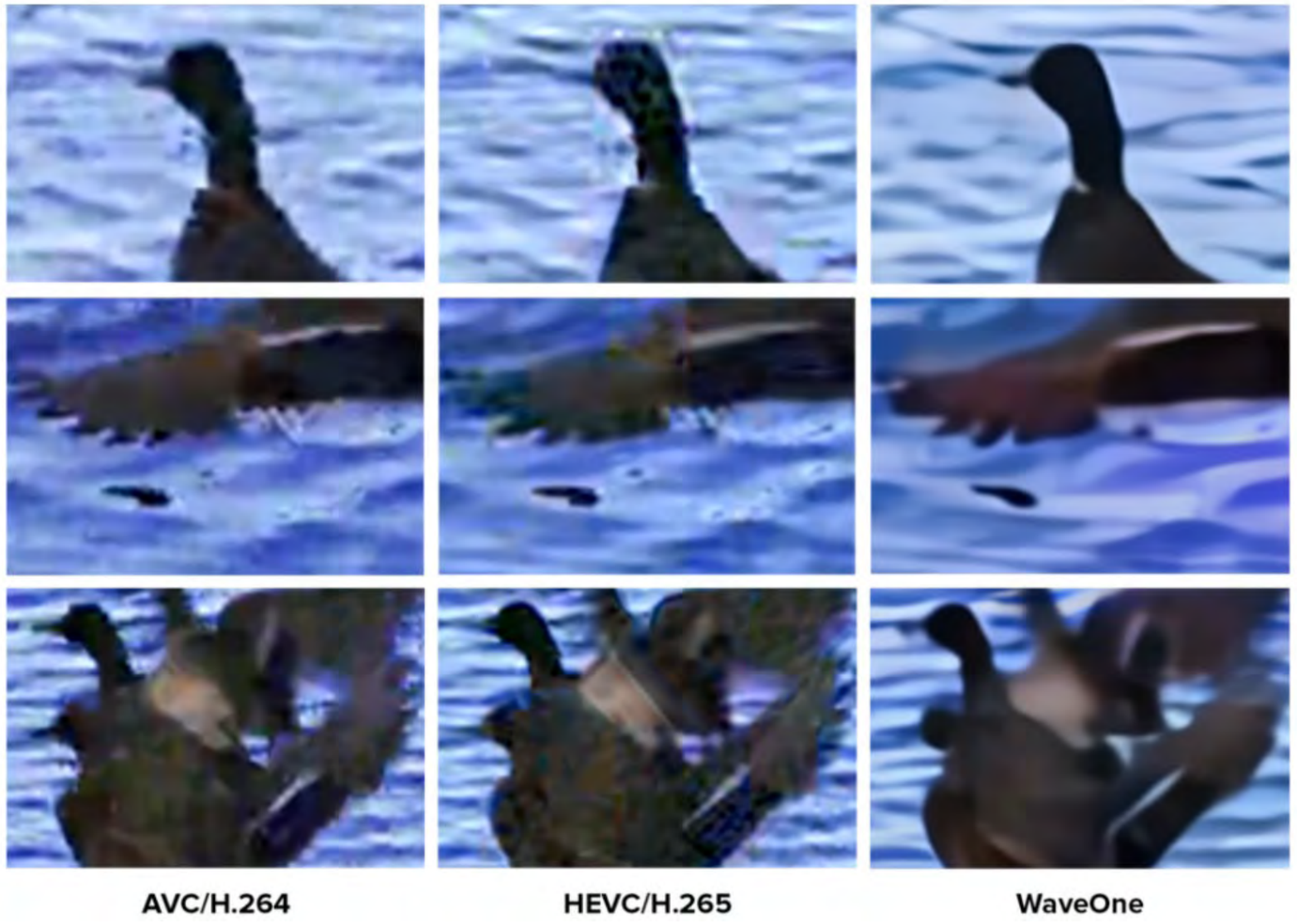
- Still Image Codec
- Auto-encoder with GAN
- Three data rates:
  - High, 3.6 bpp
  - Medium, 2.3 bpp
  - Low, 1.2 bpp
- Free and Open-source



Source: Google
Interactive HiFiC Demo at hific.github.io

WaveOne Codec

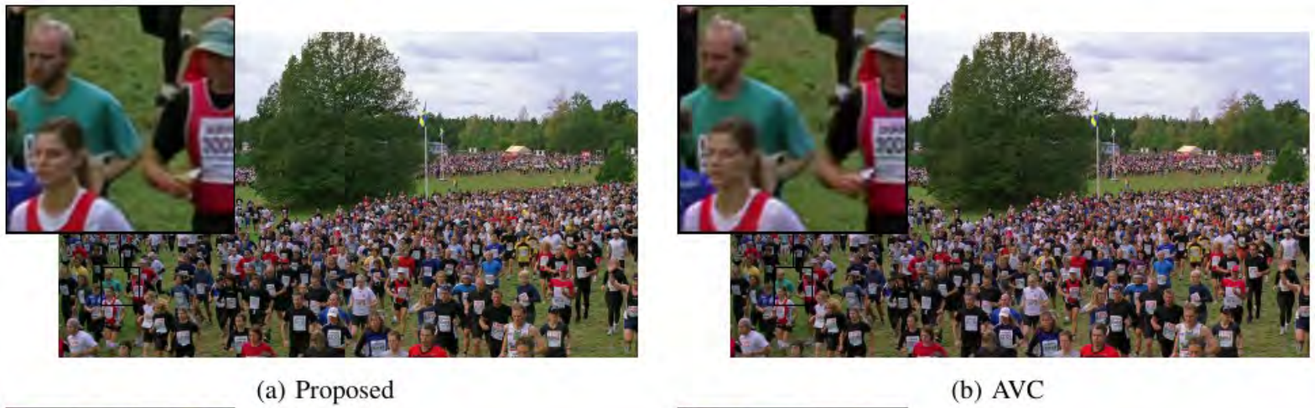
- Video Codec
- One model for various bitrates, 0.06 to 1.1 bpp
- Auto-encoder model that relies on previously decoded frame
- Proprietary codec



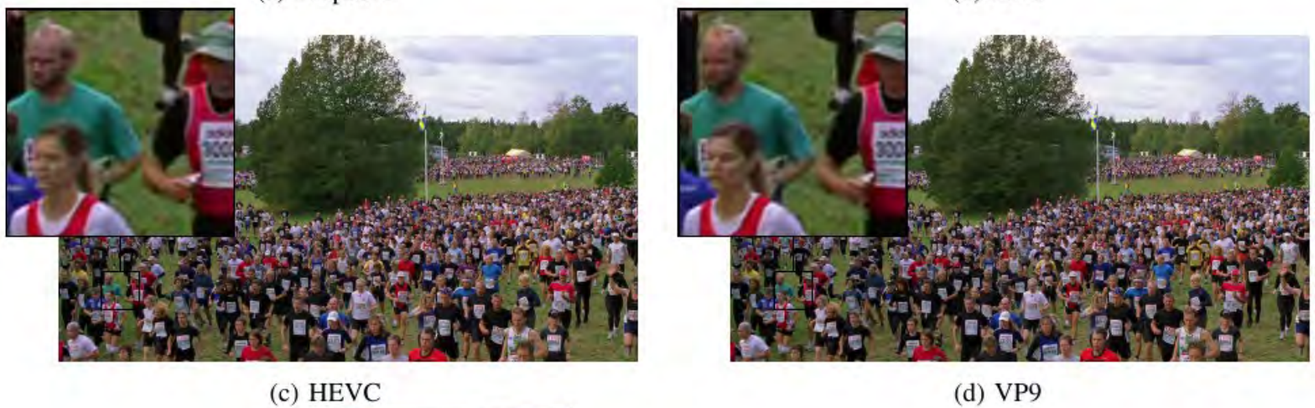
0.06 bpp, Source: WaveOne
“Learned Video Compression” by Rippel, et al., 2018

Handling Chroma Subsampling

- Current AI-based codecs use either RGB 444 or YCbCr 444 color components
- The encoder and decoder can be adjusted to handle 422 or 420 colorspace by skipping a layer for chroma
- This allows the ML model to be more efficient

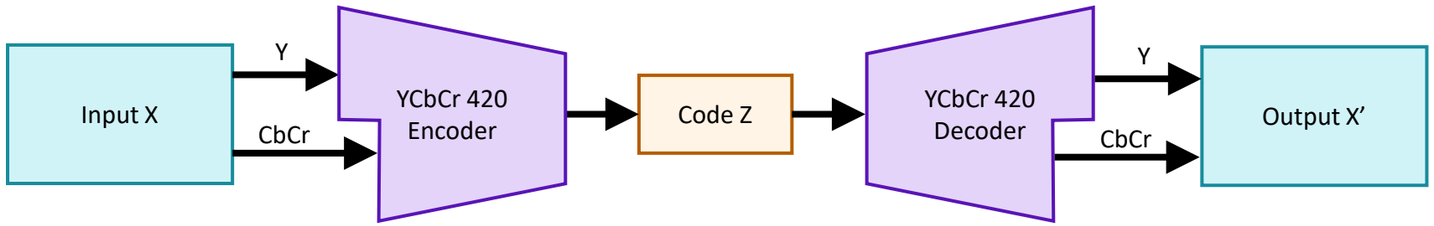


(a) Proposed (b) AVC



(c) HEVC (d) VP9

(e) Ground Truth



YCbCr 420 Auto-Encoder

0.5 bpp, Source: Xiph.org

Digital Video Metrics

Quality Metrics

- **MSE** – Mean Squared Error
- **PSNR** – Peak Signal to Noise Ratio
- **MS-SSIM** – Multi-Scale Structural Similarity Index
- **VMAF** - Video Multimethod Assessment Fusion
- Etc.

Datarate Metrics

- **bpp** – Bits Per Pixel
- **Mbps** – Megabits Per Second
- Etc.

AI-Based Video Codecs - Results

Conclusions

- The 420 auto-encoder is slightly better PSNR and MS-SSIM numbers, but a stronger improvement in VMAF, as compared to standard codecs
- The WaveOne ELF codec has a strong improvement in PSNR, and a slight improvement in MS-SSIM and VMAF as compared to HiFiC
- AI-Based Codecs are very efficient as compared to procedural codecs

Codec	Datarate (Mbps)	Datarate (bpp)	PSNR	MS-SSIM	VMAF	Speed
420 Auto-encoder	49.44	0.50	**32.00**	**98.13**	**95.85**	slow
HEVC I-Frame	49.44	0.50	31.72	97.23	86.16	fast
AVC I-Frame	49.44	0.50	30.84	96.95	83.87	fast
WaveOne ELF*	15.33	0.16	**38.20**	**99.90**	**97.56**	medium
Google HiFiC	15.33	0.16	34.12	98.20	95.41	slow

* Extrapolated from WaveOne's ELF-VC Paper, 2021

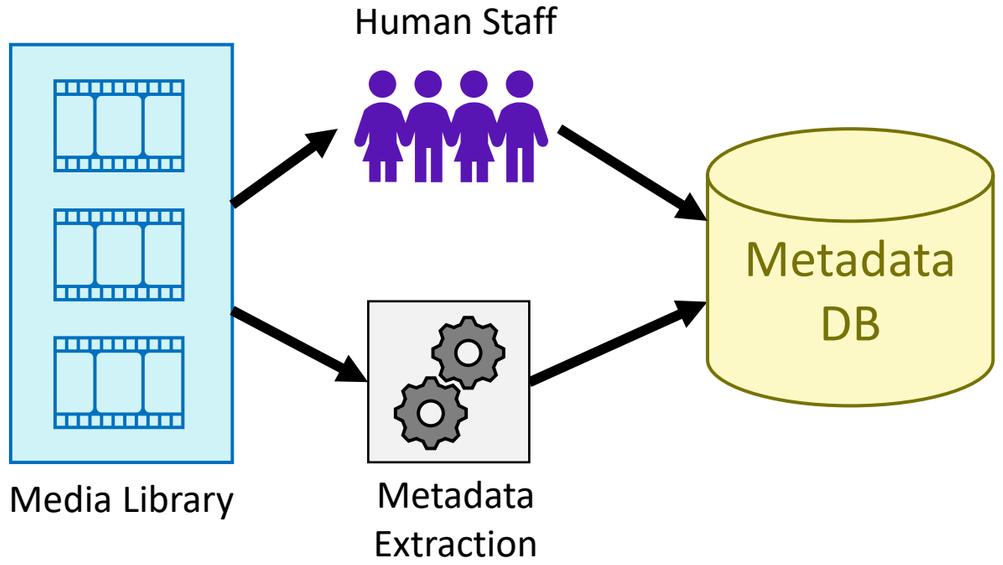
Semantic Video Search

Goal: Allow content creators to find media with simple text searches

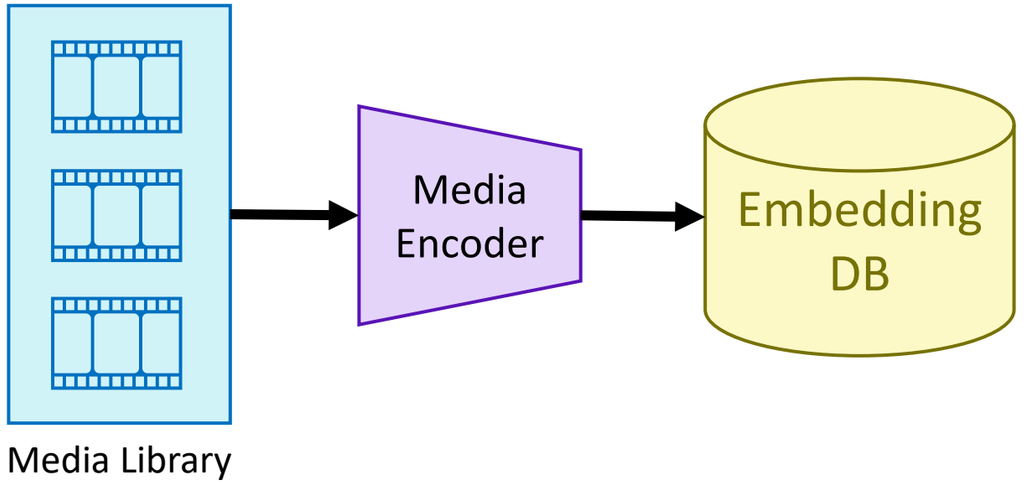
- Metadata Annotation Search vs. Free Text Search
- Semantic Search Models
  - CLIP by Open AI
  - Multilingual CLIP by Reimers and Gurevych
  - Clip4CLIP by Microsoft
- Using Elasticsearch as a database for Semantic Video Search
- Results

Metadata-based Search vs. Semantic Search

- **Metadata-based search process:**
  1. Extract metadata from media
     - Done by humans and/or automated processes
  2. Organize into taxonomy (optional)
  3. Search using keywords or known terms
  4. The system returns media that “hits”



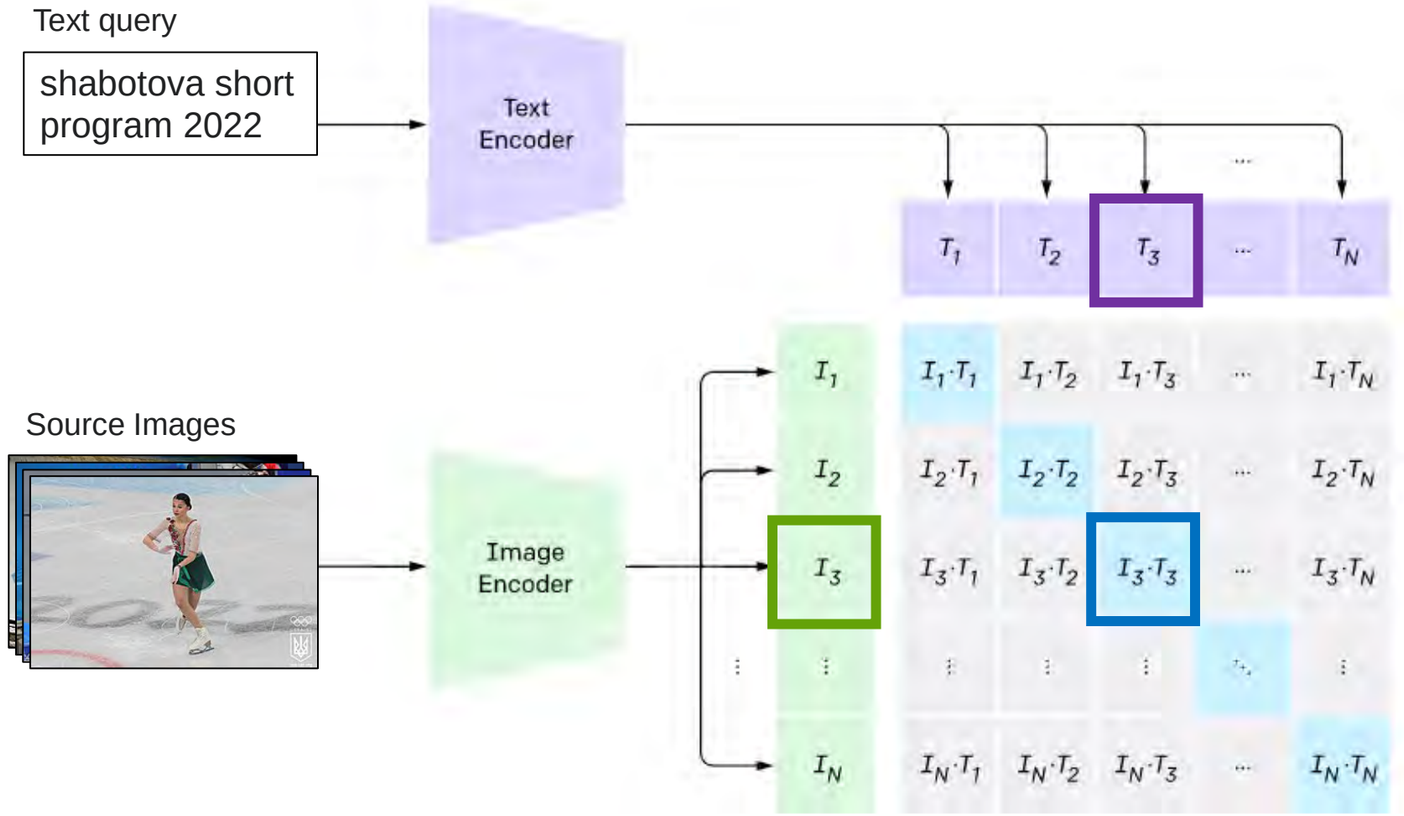
- **Semantic search process:**
  1. Create embeddings from media
     - Automated process
     - Contains semantic understanding
  2. Search with keywords or phrases
  3. The system returns closest matching media



CLIP - Image and Text Encoders

Contrastive Language-Image Pre-training

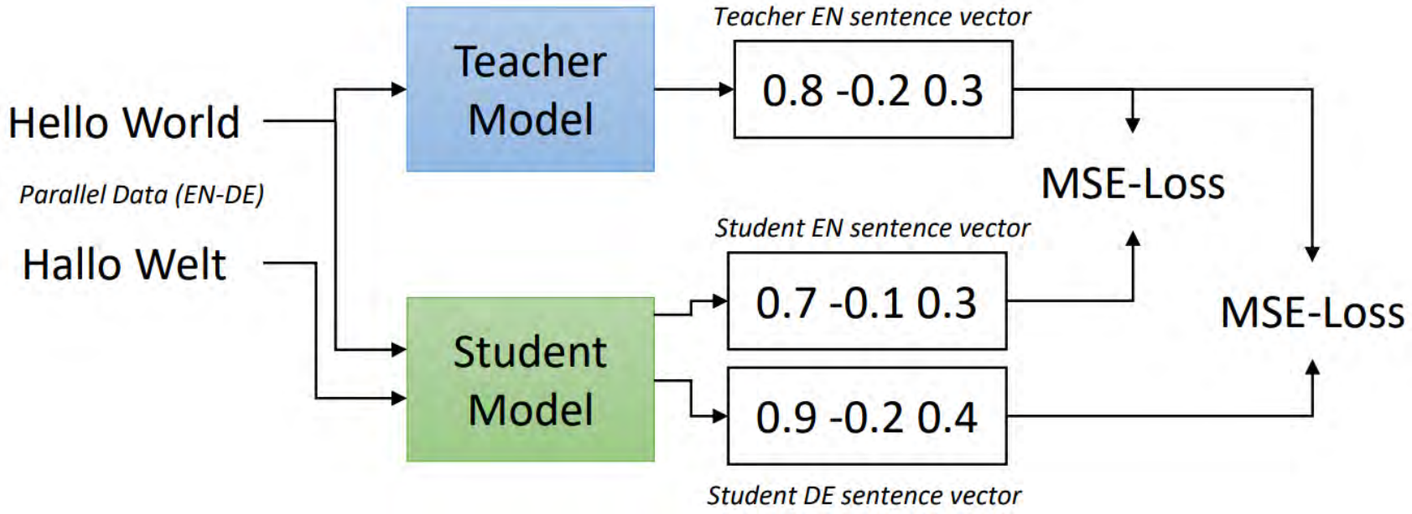
- OpenAI created the CLIP models for multi-modal semantic search
- The Text Encoder converts a phrase into an embedding, a list of 512 floating-point numbers
- The Image Encoder converts an image into a similar embedding
- We can use these models to find image features using an unstructured text query
- English Only



OpenAI's Paper: [Learning Transferable Visual Models From Natural Language Supervision]

Multilingual CLIP by Reimers and Gurevych

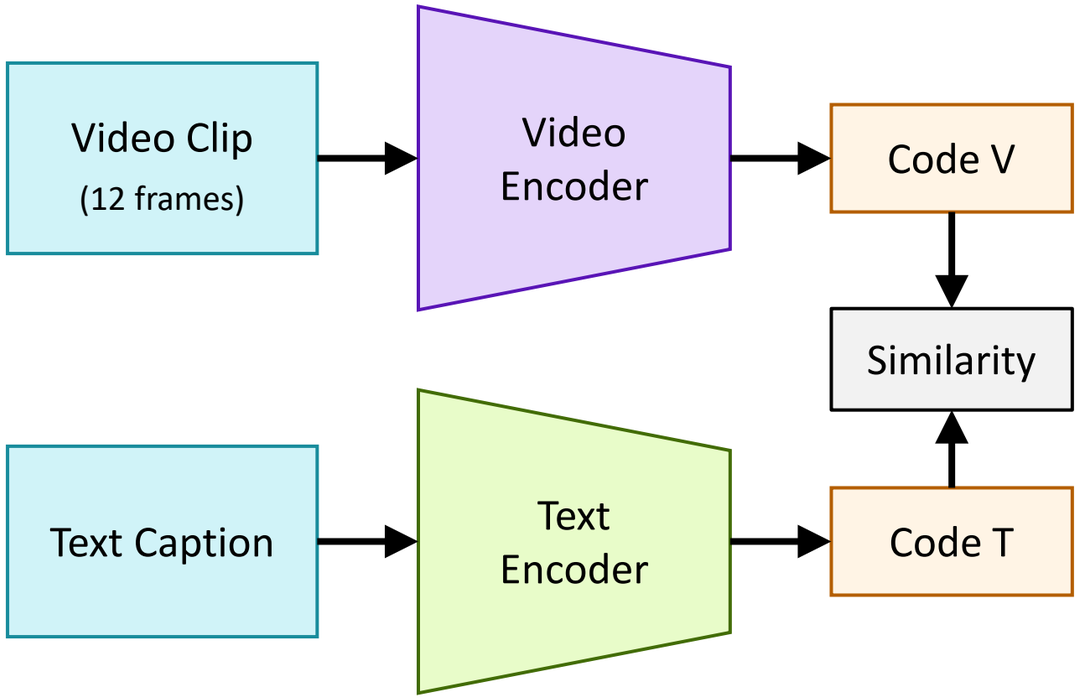
- The original CLIP model from OpenAI only supported the English language
- Nils Reimers and Iryna Gurevych from Technische Universitat Darmstadt in Germany extended CLIP to support 50+ languages
- They used Student/Teacher models to train a new Text Encoder model
- It is compatible with the original CLIP encoders
- The language is automatically detected



Paper by Reimers and Gurevych: [Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation]

Clip4CLIP by Microsoft

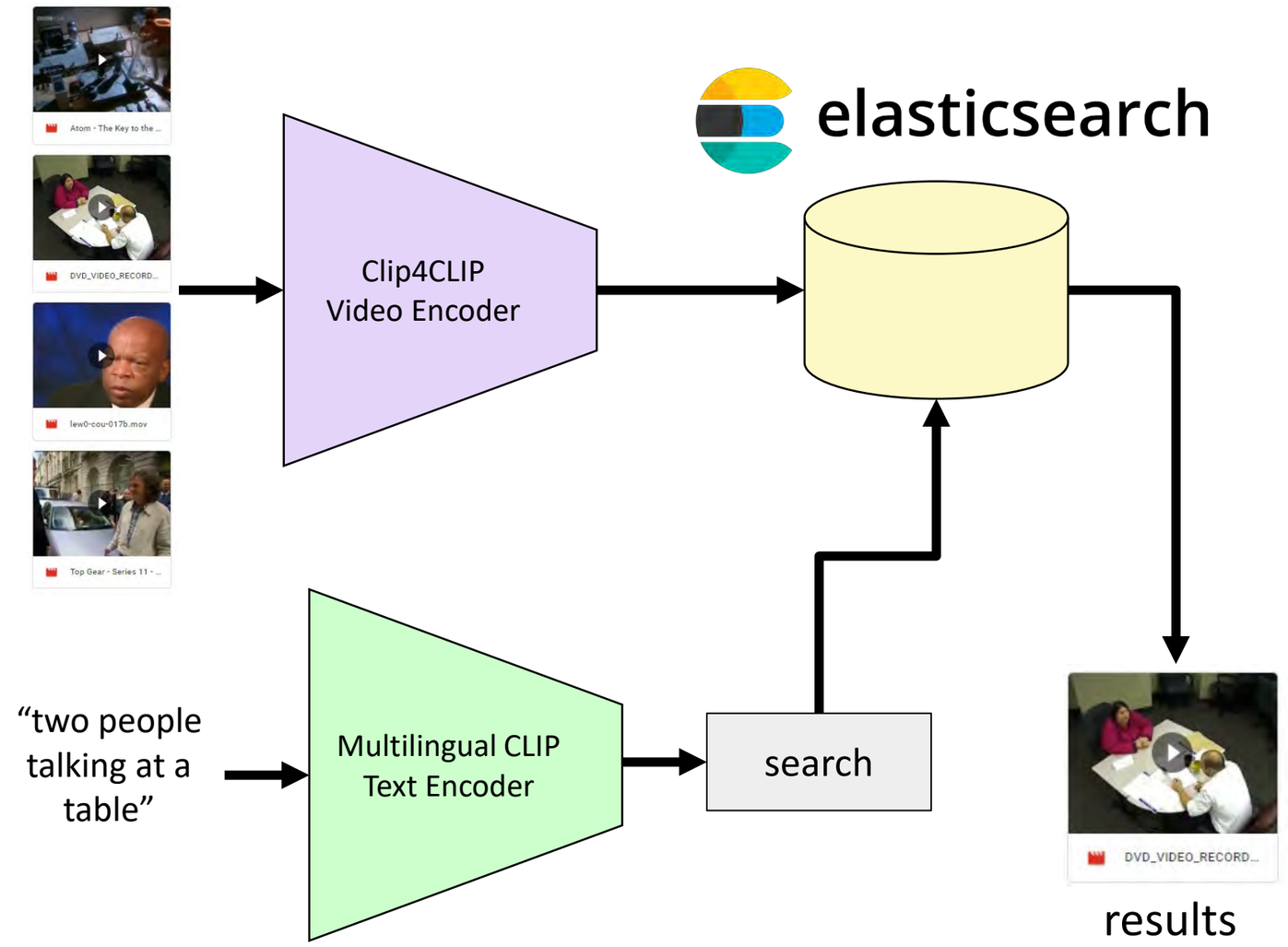
- Clip4CLIP – Retrieve video from text
  - Huaishao Luo, et al., 2021
- Video Encoder based on the Vision Transformer (ViT)
  - Alexey Dosovitskiy, et al., 2021
- Text Encoder based on OpenAI's CLIP
  - Alec Radford, et al., 2021
- The system was trained make the embedding codes for video match the embedding code for the corresponding text caption
- Video is sampled every 3.3 seconds and downsampled to 12 frames to find "actions"



Microsoft's Paper: Clip4CLIP: An Empirical Study of CLIP for End-to-End Video Clip Retrieval

Complete Semantic Video Search Solution

- Elasticsearch v7+ can be used as a database for retrieving media with indexed embeddings
- The clips are indexed using the Clip4CLIP video encoder and saved as document records in ES
- One record is created for each 3.3 seconds of video (at 30 fps)
- The clips can then be searched in any language using the Multilingual Text Encoder
- ES runs a Cosine Similarity search to find the closest matching content

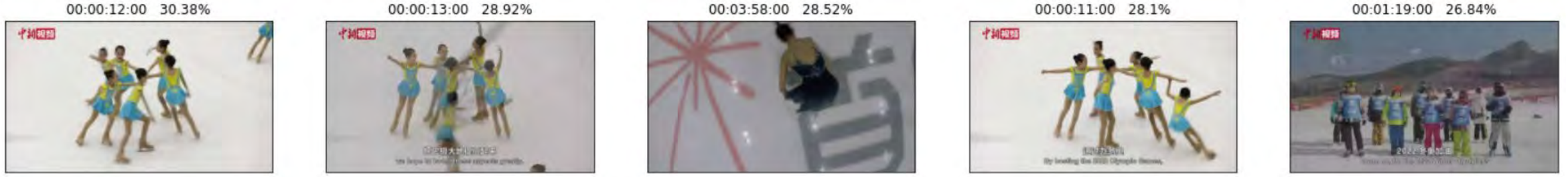


Semantic Search Results

English Search: ice dancers yellow and blue

French Search: danseurs sur glace jaune et bleu

Arabic Search: راقصات الجليد الأصفر والأزرق



Conclusions

- AI and ML can be used for many tasks in content creation
- AI-Based Video Codecs
  - Improved Quality
  - Reduced Bandwidth
- Semantic Video Search
  - Simple text searches
  - Search for actions
  - Multilingual support
  - Use Elasticsearch for DB
- Any Questions?

@Rob_Gonsalves_

medium.com/@robgon