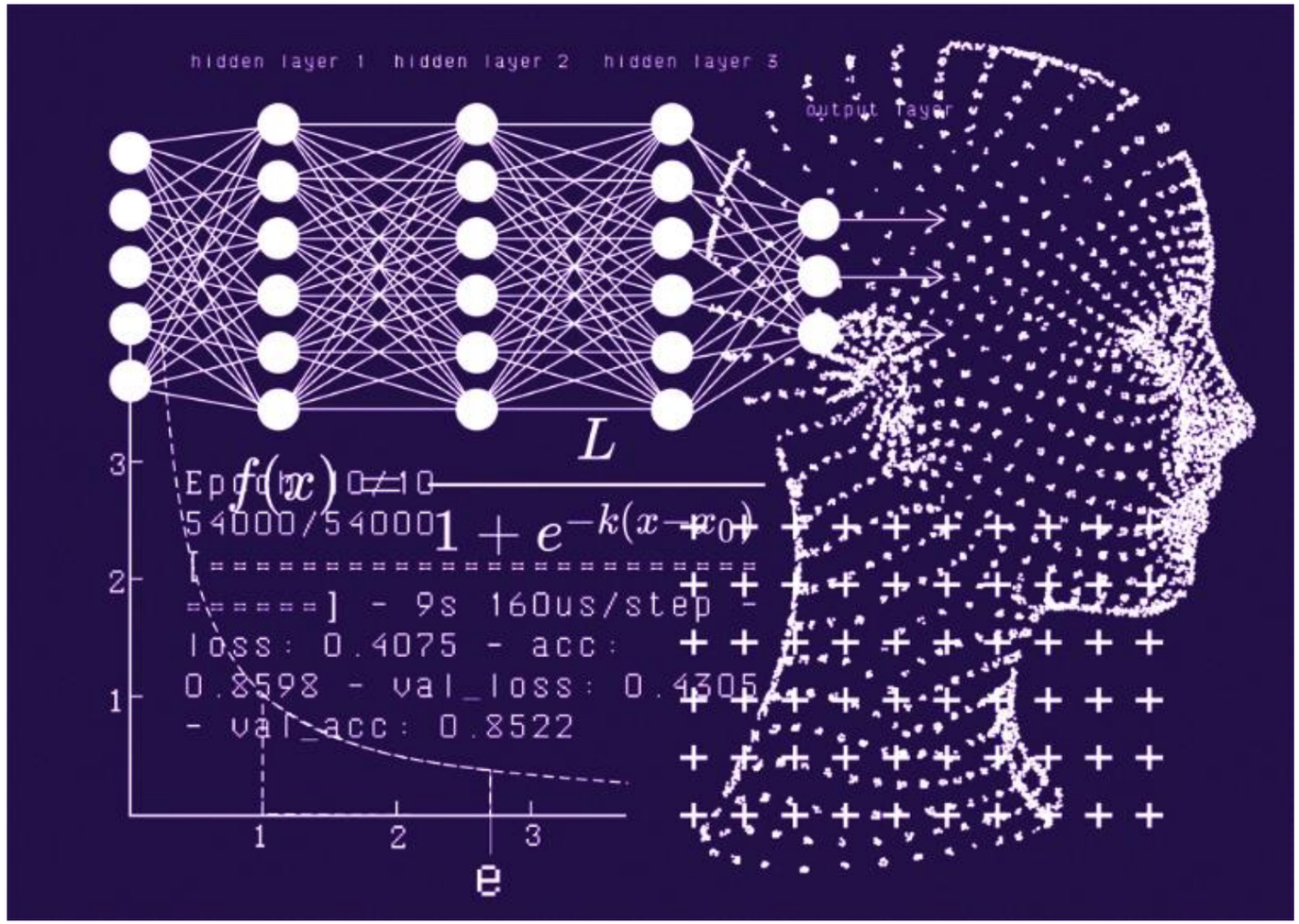
USE CASE:
TRANSFORMING MEDIA ASSET MANAGEMENT WITH AI-DRIVEN SEMANTIC CONTENT DISCOVERY

Rob Gonsalves, Avid
May 29, 2024

TRANSFORMING MEDIA ASSET MANAGEMENT WITH AI-DRIVEN SEMANTIC CONTENT DISCOVERY

- Introduction to Semantic Content Discovery
- Planning and Implementation
- Outcomes and Evaluations
- Challenges, Successes, and Next Steps
- Q&A



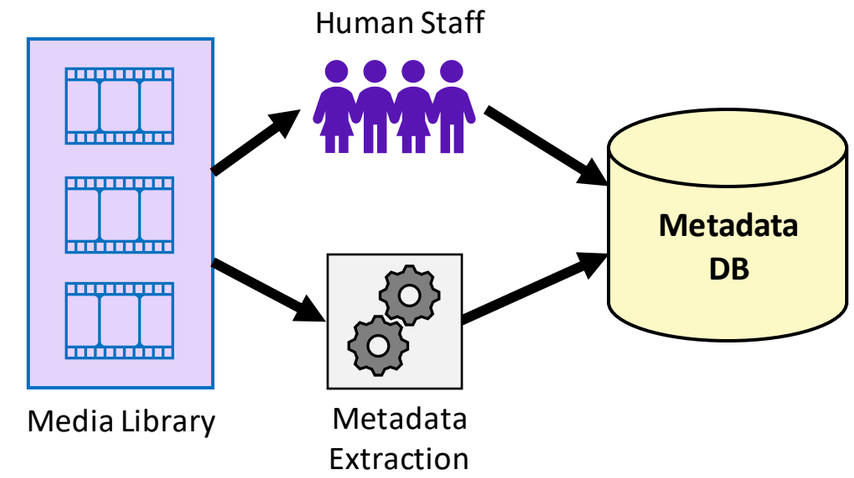
TRADITIONAL SEARCH METHODS VS. AI/ML APPROACHES

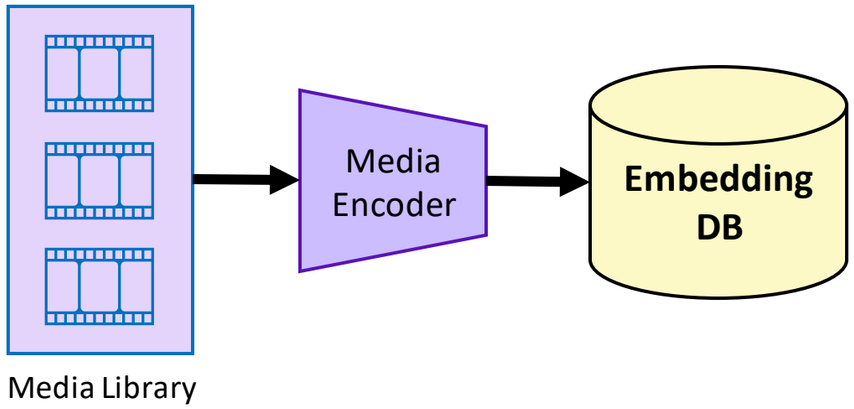
Metadata-based content discovery:

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”

AI-based semantic content discovery:

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





PLANNING AND IMPLEMENTATION

Objectives:

- Implement Semantic Content Discovery of Video Media
- Multilingual Capabilities
- Free and Open-source models

Methods:

- Compare and Evaluate Open-source Semantic Search Models
- Use Metrics like Cosine Similarity and Recall@1 for Analysis

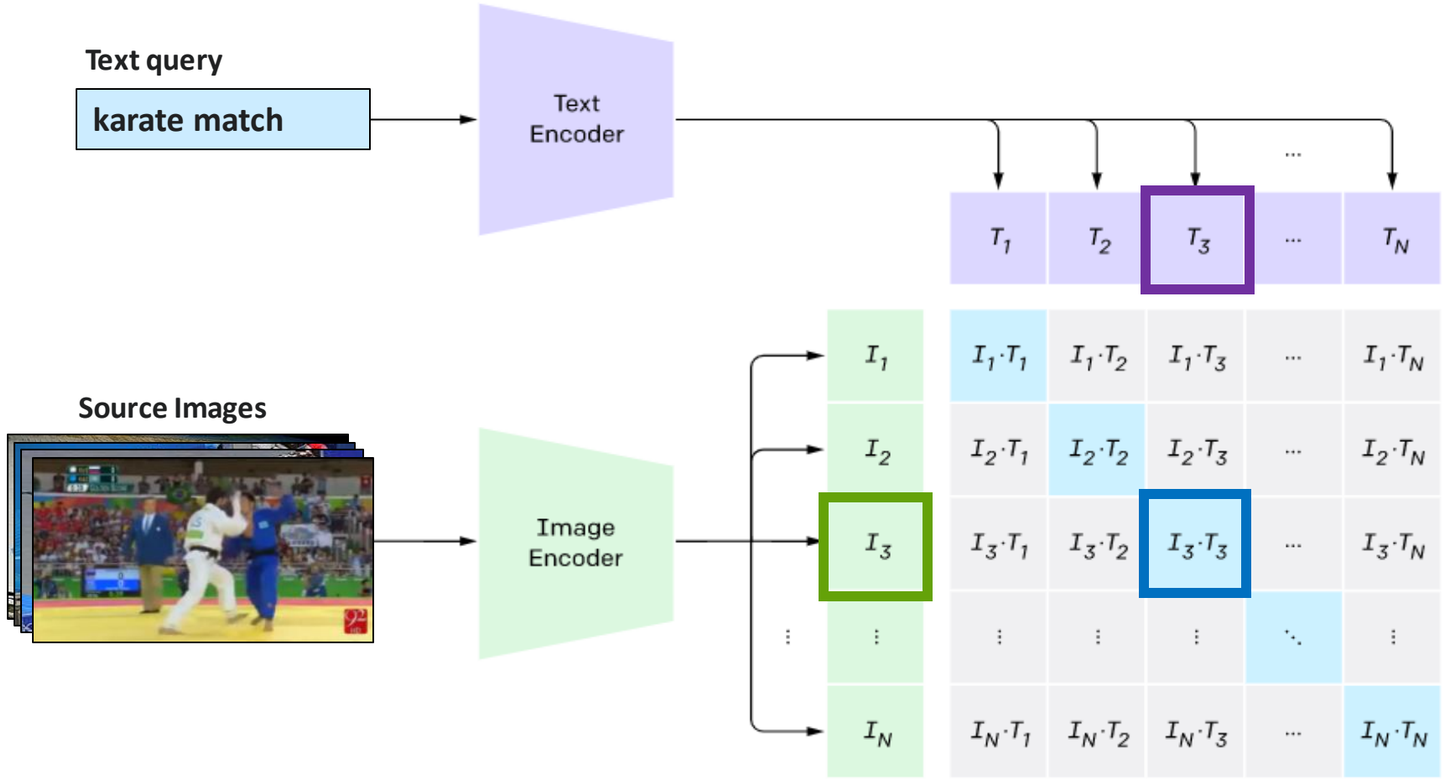
Implementation:

- Prototype in Python
- Productize in C++

CLIP - IMAGE AND TEXT ENCODERS

CONTRASTIVE LANGUAGE-IMAGE PRE-TRAINING

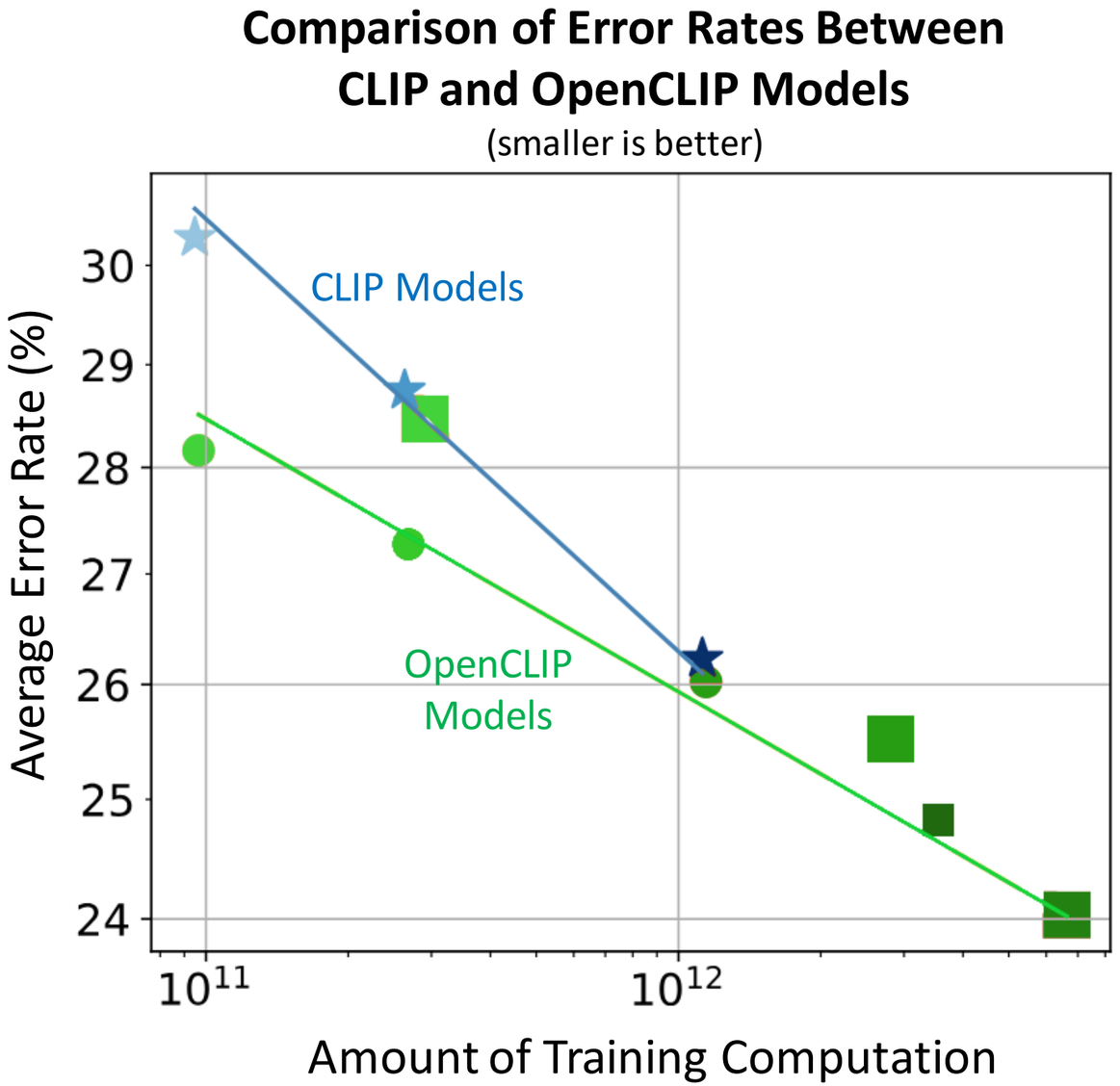
- OpenAI created the CLIP models for multi-modal semantic search
- The Text Encoder converts a phrase into an embedding, a list of 512 floating-point numbers
- The Image Encoder converts an image into a similar embedding
- We can use these models to find matching images using an unstructured text query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



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

OpenCLIP from LAION

- OpenCLIP is open-source implementation of OpenAI's CLIP.
- The OpenCLIP models were trained on up to two billion image-text pairs.
- Different scaling behaviors were observed between OpenAI and OpenCLIP models despite similar architectures and training recipes.
- Evaluated xlm-roberta-base-ViT-B-32 model using checkpoint laion5b_s13b_b90k
- The evaluation workflows, datasets, and models were open-sourced.



LAION's Paper: Reproducible scaling laws for contrastive language-image learning

SEMANTIC TEXT AND IMAGE ENCODERS

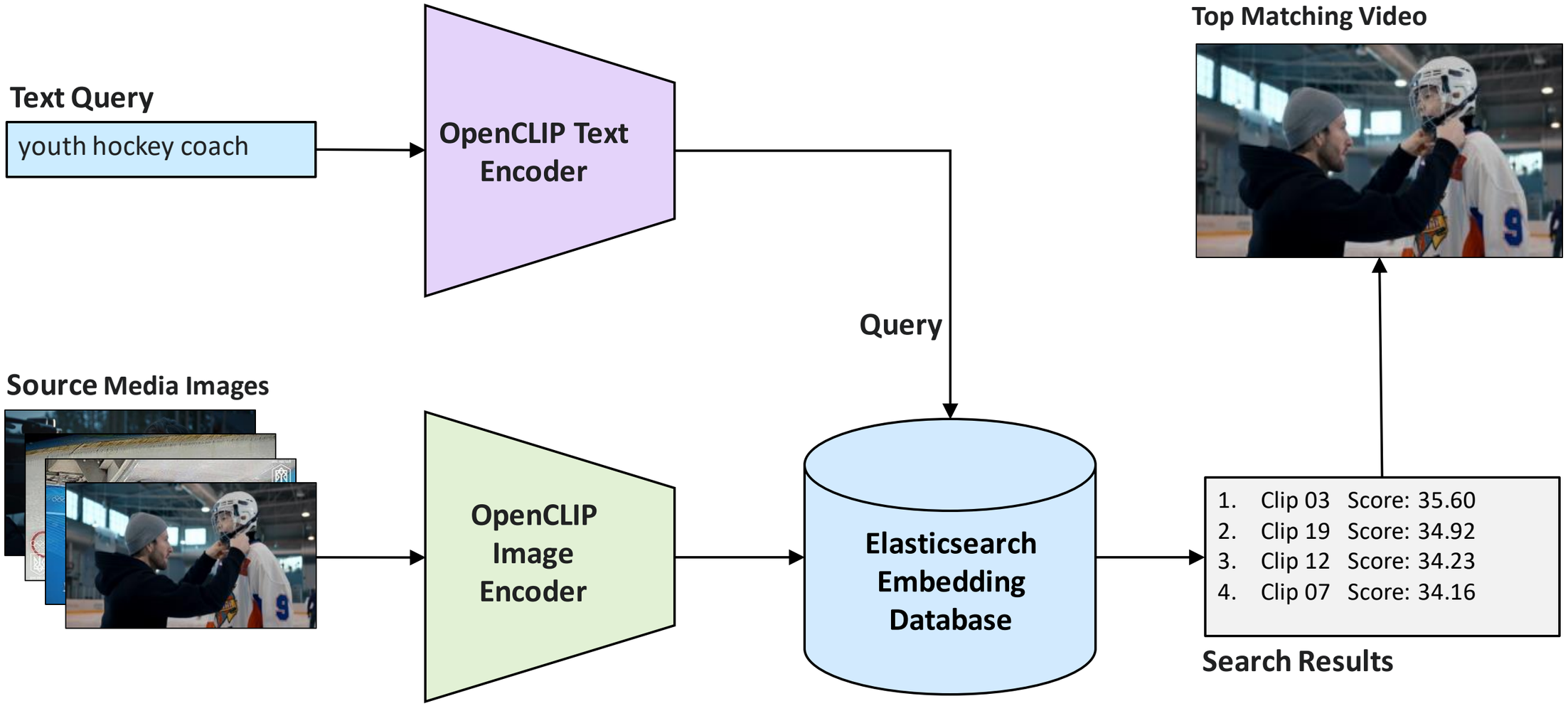


Image Source: Pexels

Planner › Assignment Landing Zone › AI in MediaCentral

AI

Claudia Roth is the presenter of MediaCentral Moments

Avid Ada Recommendations

results: 6

A-Roll

B-Roll

Score: 81.90%

Score: 80.70%

Score: 78.30%

Score: 77.40%

Score: 76.80%

Score: 76.20%

MC Moments V2 6-20_0001

01:00:00:00

01:00:00:00

01:01:42:11

Au... H... R, A ... Meta... File ... Storyb... Grap... Traco ... Trac...

L R L R

L R

S M S M

Vol 0 Vol 0

A1 A2

Stereo

+0 -4 -8 -14 -20 -26 -30 -35 -40 -60

Vol 0

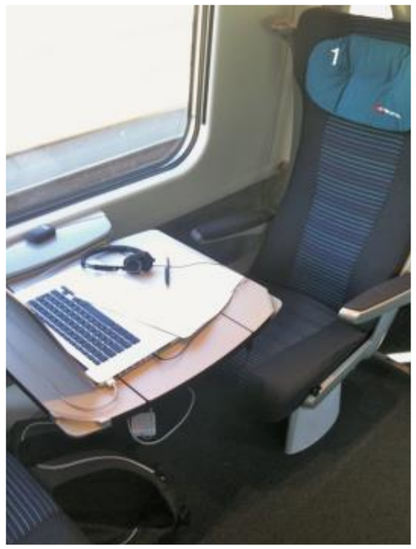
Master

OUTCOMES AND EVALUATIONS

- Search results from CLIP and OpenCLIP
- Used Google’s Crossmodal-3600 dataset for testing
- Compare results for English and multilingual searches for both models with Cosine Similarity and Recall@1 metrics
- Compare image encoding times for both models

Results from CLIP, search term: “A laptop on a train.”

Score: 0.3582



Score: 0.2998

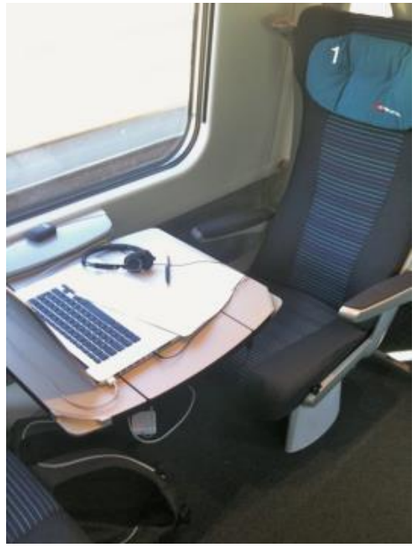
Score: 0.2908

Score: 0.2690

Score: 0.2673

Results from OpenCLIP, search term: “A laptop on a train.”

Score: 0.3669



Score: 0.3069

Score: 0.2964

Score: 0.2778

Score: 0.2659

Results from CLIP, search term: “Two people talking in a living room.”

Score: 0.2927

Score: 0.2920

Score: 0.2808

Score: 0.2681

Score: 0.2644

Results from OpenCLIP, search term: “Two people talking in a living room.”

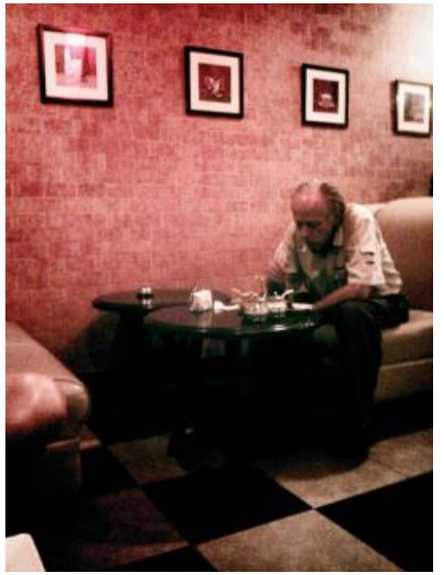
Score: 0.2656

Score: 0.1886

Score: 0.1874

Score: 0.1860

Score: 0.1743



CLIP - English

Median of matches (diagonal): 0.309
Median of misses (off-diagonal): 0.149
Median Match-to-Miss Delta: 0.160

A young man smiling and posing.	0.28	0.17	0.16	0.17	0.17	0.20	0.13	0.19
A group of people having fun in rafting on the water.	0.17	0.30	0.13	0.14	0.09	0.18	0.16	0.18
A macro shot of a candy floss cover.	0.16	0.13	0.29	0.21	0.10	0.07	0.13	0.19
A macro shot of white and pink flowers with a background of white clouds in a blue sky.	0.13	0.08	0.15	0.34	0.02	0.08	0.09	0.10
An interior view of an old abandoned factory.	0.15	0.15	0.16	0.16	0.34	0.21	0.16	0.13
An old house in the village through the trees.	0.14	0.12	0.07	0.17	0.13	0.26	0.09	0.09
The log sheet.	0.19	0.22	0.19	0.17	0.20	0.20	0.32	0.19
Chinese steamed dim sum in various colors placed in a basket.	0.16	0.14	0.13	0.13	0.09	0.11	0.13	0.34

OpenCLIP - English

Median of matches (diagonal): 0.249
Median of misses (off-diagonal): 0.013
Median Match-to-Miss Delta: 0.236

A young man smiling and posing.	0.22	0.00	0.01	-0.03	0.01	0.01	-0.02	-0.03
A group of people having fun in rafting on the water.	0.02	0.23	0.00	-0.02	-0.01	0.05	0.01	0.02
A macro shot of a candy floss cover.	0.09	0.06	0.33	0.15	0.07	0.04	0.02	0.13
A macro shot of white and pink flowers with a background of white clouds in a blue sky.	0.01	-0.09	0.03	0.26	-0.04	-0.02	0.02	0.01
An interior view of an old abandoned factory.	0.03	-0.02	0.02	0.00	0.29	0.08	0.03	-0.01
An old house in the village through the trees.	0.00	-0.04	-0.03	0.05	0.03	0.20	-0.06	-0.04
The log sheet.	0.08	0.13	0.10	0.09	0.15	0.13	0.24	0.08
Chinese steamed dim sum in various colors placed in a basket.	-0.02	-0.02	0.01	-0.01	-0.03	-0.01	0.05	0.26

CLIP - Multilingual

Median of matches (diagonal): 0.227
Median of misses (off-diagonal): 0.175
Median Match-to-Miss Delta: 0.052

(English) A young man smiling and posing.	0.28	0.17	0.16	0.17	0.17	0.20	0.13	0.19
(French) 7 personnes sur une boué jaune et orange faisant du rafting en vertical dans l'eau	0.14	0.30	0.12	0.18	0.14	0.17	0.18	0.17
(Korean) 타투가 들어있는 사탕 케이스 껍질	0.19	0.16	0.19	0.19	0.20	0.17	0.23	0.20
(Ukrainian) Рожевий квіт гілки фруктового дерева на тлі голубого з хмарами неба	0.20	0.19	0.18	0.21	0.21	0.19	0.24	0.17
(Persian) نمایی از داخل یک ساختمان قدیمی و مخروبه و نورکه از لابه لای پنجره ها به داخل تابیده است	0.18	0.17	0.15	0.22	0.18	0.16	0.20	0.18
(Bengali) চারদিকে সবুজ গাছপালা আছে এবং মাঝে মুরগী দেখো যাচ্ছে	0.16	0.20	0.15	0.20	0.17	0.25	0.18	0.18
(Greek) Φόρμα πληροφοριών	0.21	0.18	0.17	0.22	0.20	0.17	0.21	0.16
(Indonesian) Salah satu dim sum yang dinamakan xiaolongbao disajikan dengan beragam varian warna	0.15	0.13	0.13	0.16	0.17	0.13	0.19	0.31

OpenCLIP - Multilingual

Median of matches (diagonal): 0.235
Median of misses (off-diagonal): 0.045
Median Match-to-Miss Delta: 0.190

(English) A young man smiling and posing.	0.22	0.01	0.01	-0.03	0.02	0.02	-0.01	-0.03
(French) 7 personnes sur une boué jaune et orange faisant du rafting en vertical dans l'eau	-0.00	0.27	-0.02	-0.04	-0.01	0.04	0.04	0.03
(Korean) 타투가 들어있는 사탕 케이스 껍질	0.10	0.05	0.23	0.08	0.09	0.06	0.05	0.16
(Ukrainian) Рожевий квіт гілки фруктового дерева на тлі голубого з хмарами неба	0.03	-0.04	0.08	0.26	-0.00	0.04	0.01	0.00
(Persian) نمایی از داخل یک ساختمان قدیمی و مخروبه و نورکه از لابه لای پنجره ها به داخل تابیده است	0.07	0.04	0.01	0.03	0.24	0.10	0.04	0.05
(Bengali) চারদিকে সবুজ গাছপালা আছে এবং মাঝে মুরগী দেখো যাচ্ছে	0.08	0.10	0.08	0.14	0.09	0.15	0.11	0.10
(Greek) Φόρμα πληροφοριών	0.12	0.09	0.09	0.08	0.12	0.07	0.15	0.06
(Indonesian) Salah satu dim sum yang dinamakan xiaolongbao disajikan dengan beragam varian warna	-0.00	0.03	0.05	0.04	0.01	0.03	0.09	0.30

Results from CLIP, Search term: “**거실에서 이야기를 나누는 두 사람.**”

“Two people talking in a living room.” in Korean

Score: 0.2717

Score: 0.2698

Score: 0.2637

Score: 0.2598

Score: 0.2593

Results from OpenCLIP, Search term: “**거실에서 이야기를 나누는 두 사람.**”

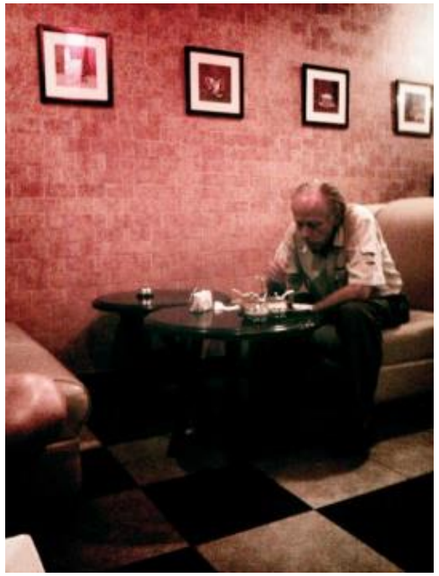
“Two people talking in a living room.” in Korean

Score: 0.2771

Score: 0.2100

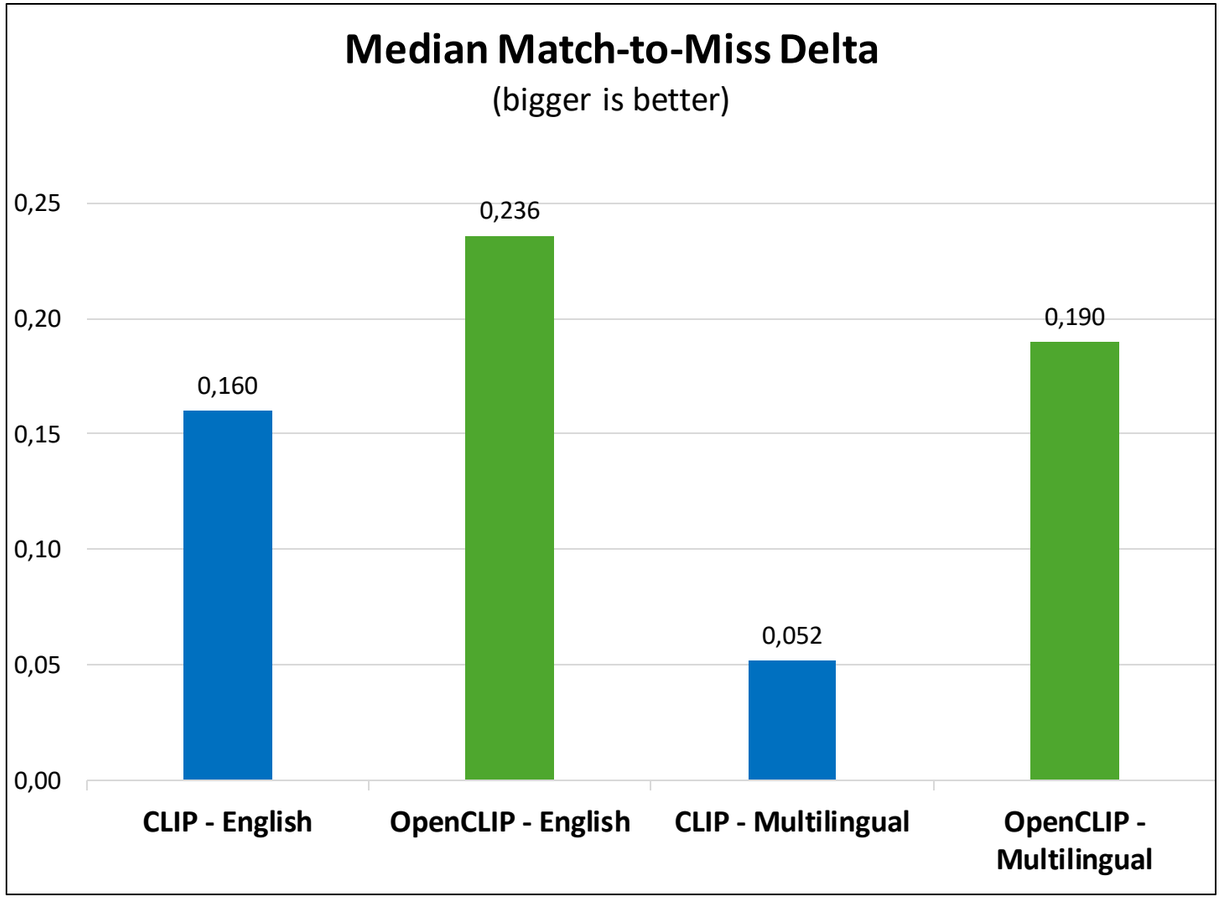
Score: 0.2067

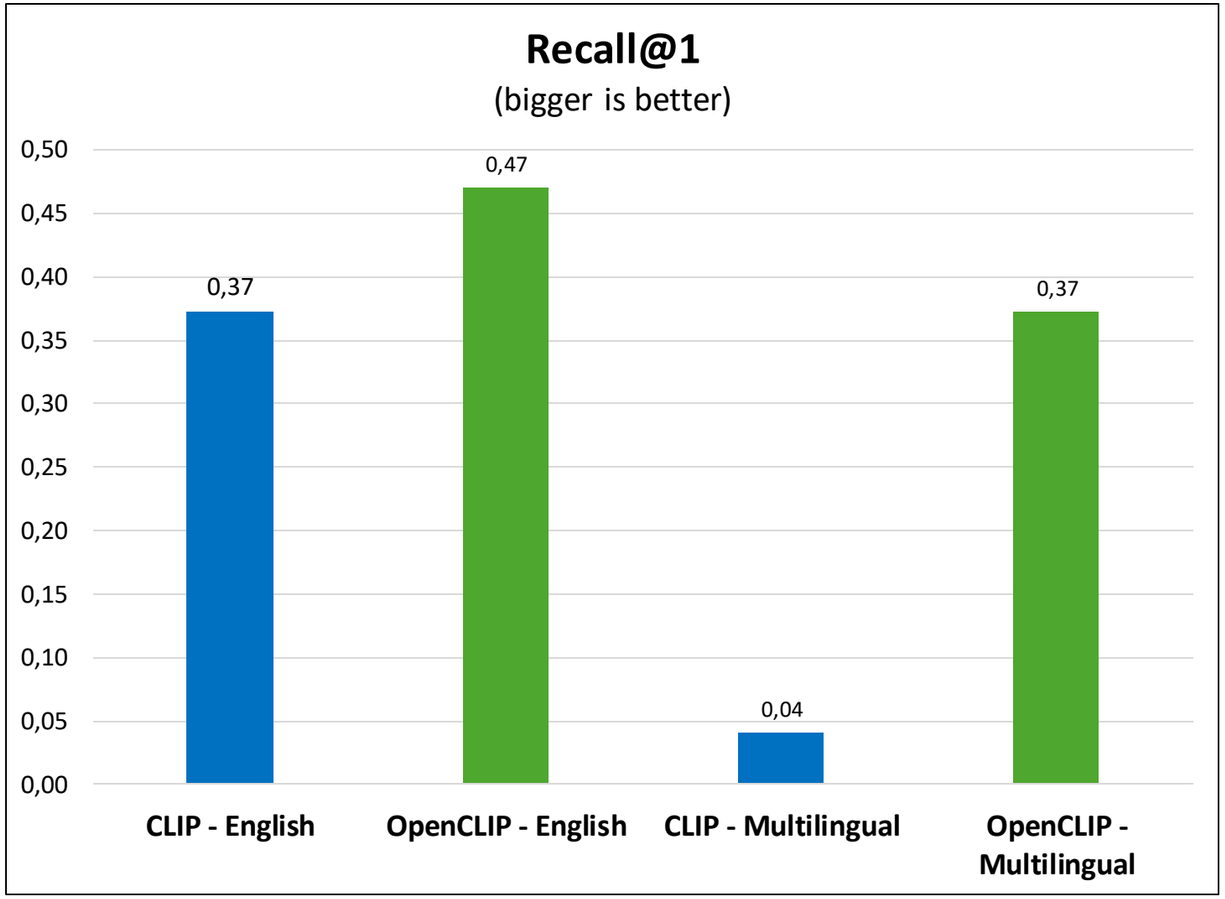
Score: 0.1974



Score: 0.1962

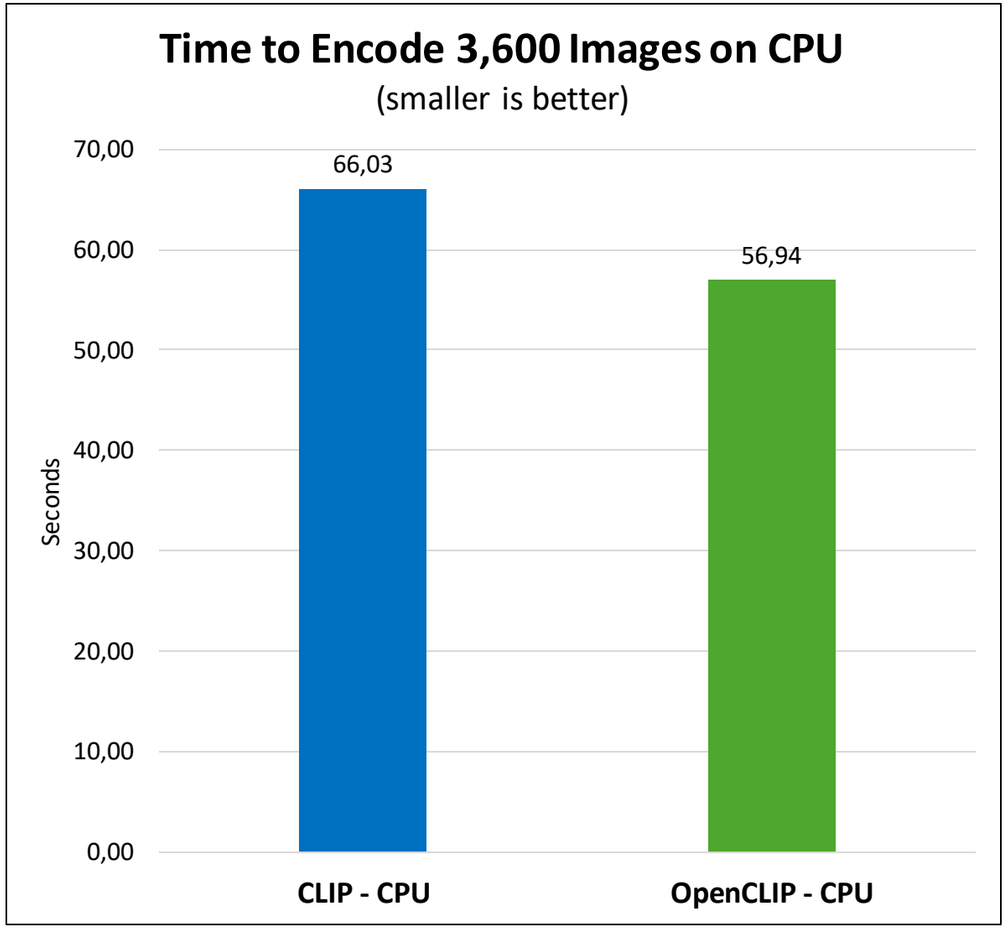
Accuracy Results of CLIP and OpenCLIP

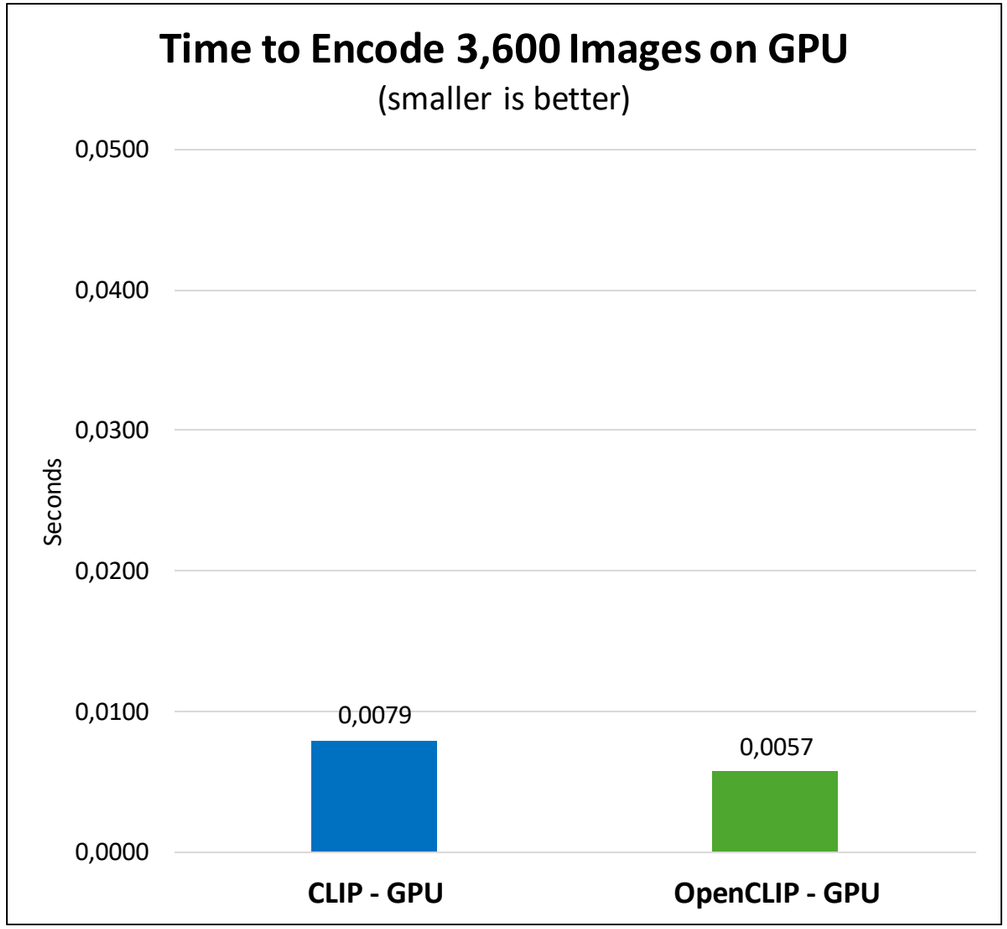




Dataset: Google's Crossmodal-3600
Recall@1 results from LAION

Performance Results of CLIP and OpenCLIP





Dataset: Google's Crossmodal-3600
CPU System: Lenovo ThinkPad P16, Core i9, 24 Cores, 2.3 GHz, 64 GB RAM
GPU System: NVIDIA RTX A5500, 16 GB RAM

CHALLENGES AND NEXT STEPS

Challenges:

- Scalability and performance optimization for large volumes of media data
- Ensuring high accuracy across languages
- Extending search to other media types like video and audio

Next Steps:

- Optimizing encoding speeds with model quantization
- Further fine-tuning models for specific use cases

CONCLUSION

- Semantic Content Discovery is a method of searching for and retrieving media files
- A database can be used to store the resulting embeddings
- Search using unstructured text queries
- LAION’s OpenCLIP outperforms OpenAI’s CLIP on many criteria
- OpenCLIP works well with multiple languages

QR code for Avid Ada

robgon-art

@Rob_Gonsalves_

medium.com/@robgon