Similarity Detection in Online Integrity

Fighting abusive content with algorithms

Alberto Massidda,
Production Engineer

Meta

Outline

1. The problem
2. The role of Automation and Similarity Detection
3. The current technology for images: vector search
4. The embeddings: PhotoDNA, PDQ/VideoPDQ, SSN++
5. The current platform: ThreatExchange
6. The current FOSS offering: Hasher-Matcher-Actioner

The problem

- Any big platform bears the responsibility to ensure it is a safe place to surf.
- Nearly 3B users. Vast majority follow rules, some fringe bad actor always present.
- Issues like: Child Exploitation, Non Consensual Intimate Imagery (read, revenge porn), Adult Sexual Exploitation, Terrorism, Violence, etc.

The problem

- Q2’22 38M Adult Sexual Exploitation taken down; 0.04% of viewed content.
  - 97.2% proactively taken off. 500k restored.

- Sheer volume of content reviewed daily requires both automation and human review to ensure accuracy and consistency

The role of Automation and Similarity Detection

As any other actor, Meta employs automation to:

- Scale
- Consistently repeat decisions of human reviewers

We tie:

- Content to Decisions
- Decisions to Actions

We do that for video, images and text.

This presentation will be mostly about images.

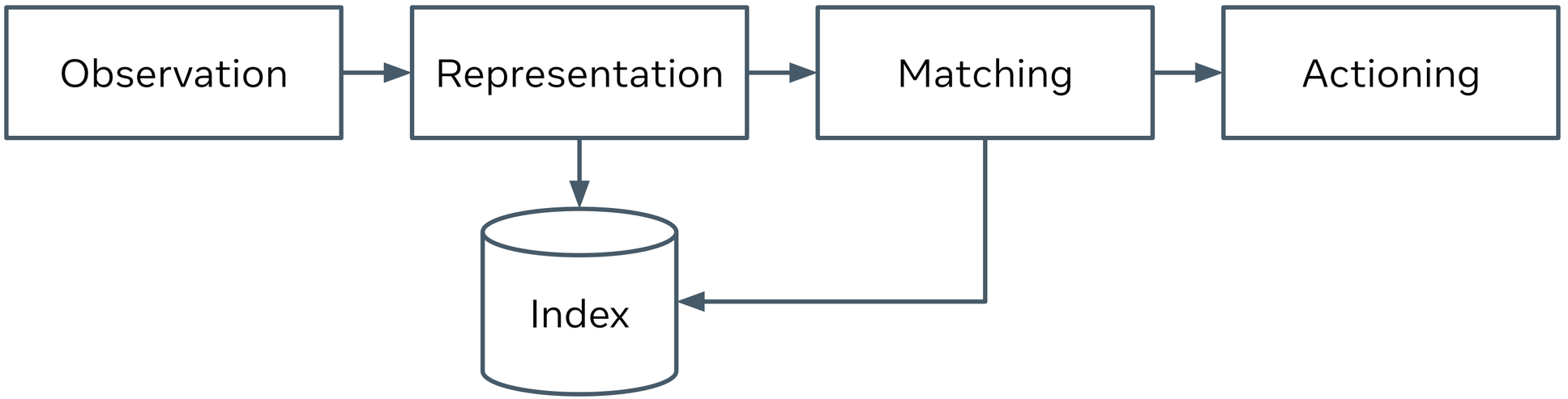
Similarity Detection in Images as Vector Search

Crypto hashing is not resistant to resize, rotation, whitening, 1 pixel alteration.

Local hashing allows for similarity measurement: **turn an image into a vector and perform vector search**.

A base SD architecture

1. Observation: an image has been generated (usually, a push event)
2. Representation: hashing the image to a compact representation
3. Matching: searching the index
4. Actioning: what do you what to do with it



Similarity Detection in Images as Vector Search

Crypto hashing is not resistant to resize, rotation, whitening, 1 pixel alteration.

Local hashing allows for similarity measurement: **turn an image into a vector and perform vector search**.

FAISS (github.com/facebookresearch/faiss) is a library for similarity search of dense vectors.
C++ version of Lucene, on CUDA steroids. Python bindings available.

Similarity Detection in Images as Vector Search

Crypto hashing is not resistant to resize, rotation, whitening, 1 pixel alteration.

Local hashing allows for similarity measurement: **turn an image into a vector and perform vector search**.

FAISS (github.com/facebookresearch/faiss) is a library for similarity search of dense vectors.
C++ version of Lucene, on CUDA steroids. Python bindings available.

We mostly refer to perceptual hashing: captures the visual similarities.

Do we really need ConvNets for that?

2009: Microsoft invents PhotoDNA

PhotoDNA is the first notable algo employed in fight against exploitive imagery of children.

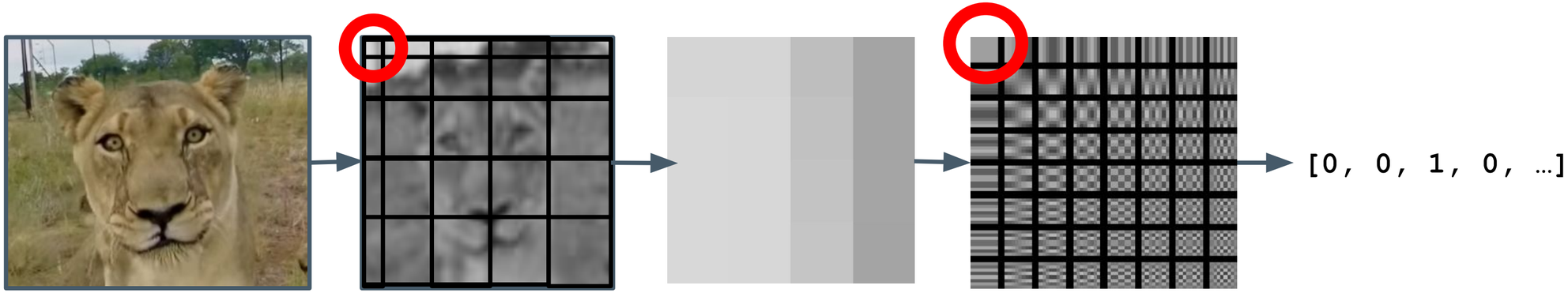
- Computes a hash of 144 uint8.
- It’s proprietary, so its details cannot be disclosed.
- It’s only for Child Exploitation Imagery. Can’t be used for other content type.

Microsoft donated PhotoDNA to the National Center for Missing & Exploited Children (NCMEC) and shares it with any organization fighting child abuse.

2019: Facebook releases PDQ

- A Perceptual algorithm utilising a Discrete Cosine Transform and outputting a Quality metric.
- 256 bit hash, uses Hamming distance.
- Very fast, compute time negligible compared to disk read.
- Can tolerate minimum adversariality.
- Used in StopNCII.org

PDQ



Hashing is:

1. (optional) Scale down to 512 x 512.
2. Compute luminance of each pixel.
3. Downsample to 64 x 64 using a blur filter to get the most significative value.
4. Divide the image in 16 x 16 boxes, each one 4 x 4 pixels.
5. Calculate a DCT of each box:
   if the number is above the median of each box, it's 1. Otherwise it's 0.
   You get 16 x 16 = 256 bits vector.

DCT provides a spectral-hashing property: identifies what contributes more or less to the image.

Hashing space is 2^{128}.

Searching is: do a vector search.

Video hashing: TMK + PDQF

TMK (for Temporal Match Kernel) is a video-similarity-detection algorithm.

It produces fixed-length video hashes.

Hashing is:

1. Resample a video to 15 fps.
2. Compute PDQ-f (PDQ without 0-1 quantization, so it’s floats) for every frame
3. Compute average of descriptors within various periods over cos and sin (keeps time signature).

Video hashing: TMK + PDQF

TMK (for Temporal Match Kernel) is a video-similarity-detection algorithm.

It produces fixed-length video hashes.

Hashing is:

1. Resample a video to 15 fps.
2. Compute PDQ-f (PDQ without 0-1 quantization, so it's floats) for every frame
3. Compute average of descriptors within various periods over cos and sin (keeps time signature).

Searching is:

1. Compare vector 0, the average of all descriptors ("level-1", loses time references, faster)
2. Compare all other vectors at different periods ("level-2")

Hashing is slow.

VideoMD5

I lied: we use crypto hashes for videos.

- Take MD5 of video and find exact copies.
- Can be done with vector search if we use the bytes.
- Used in StopNCII.org

2022: Facebook releases VideoPDQ

Hashing is:

- Hash every frame to a PDQ hash and pack the list. That is a VideoPDQ hash, of variable length.

Searching is:

- Find one or more matching frame(s);
- Pull all the frames from the query video and from the candidate to do a pairwise comparison. Above a certain consecutive threshold, we have a match.

Threat Exchange platform

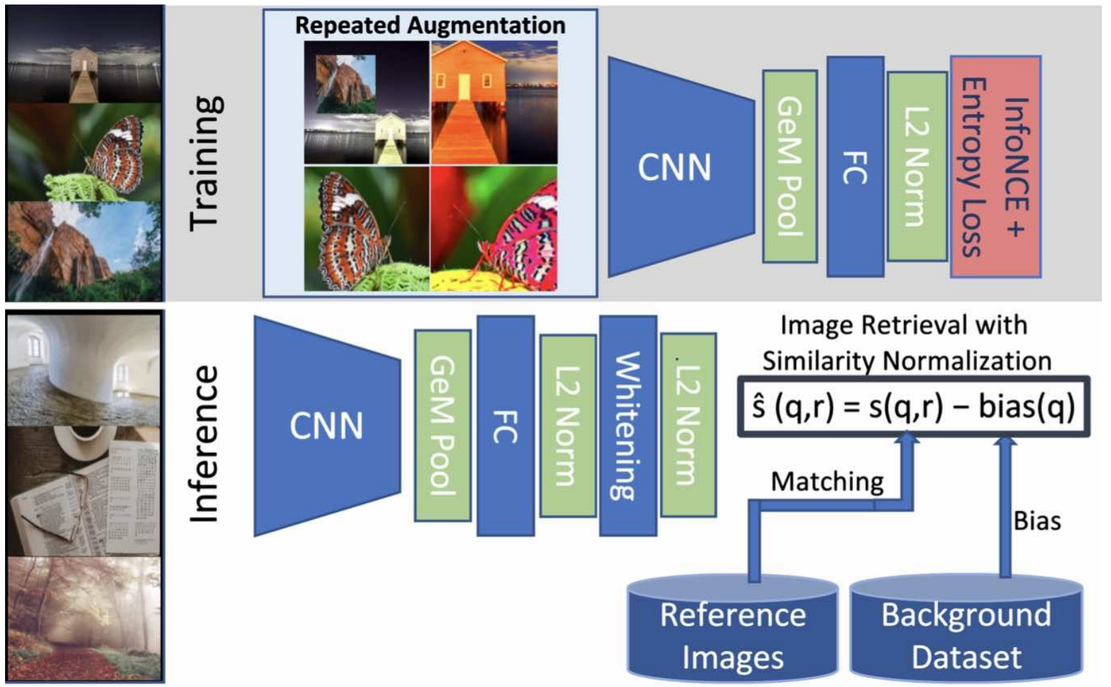
NCMEC shares PDNA hashes with all companies asking for them.

Meta's *Internet Safety Engineering* team builds and operates a service that allows companies to upload (PDQ hashes) seeds to a graph and share them with other actors.

- Exposes ReST APIs to access and POST new data.
- Has multilang clients.
- Uses PDQ.
- Users can download data.

github.com/facebook/ThreatExchange

2020: SimSearchNet++
2022: SSCD



State of the art.

- Pytorch based. Models and code available.
- ResNet-50 CNN, based on R-MAC vocabularies.
  - Regional MAC (Maximum Activation of Convolutions):
    region where there is the max pooling of activations across channels.
    Interesting regions have high activations. Use R-MAC as words in a *cosine-similarity* search.
- Self-supervised: Trained to recognize augmented input to original input.
  - Highly resistant to adversarial manipulation.

github.com/facebookresearch/sscd-copy-detection

Image Similarity Challenge

Determine whether a query image is a modified copy of any image in a reference corpus of size 1 million.

https://sites.google.com/view/isc2021

Meta AI Video Similarity Challenge

- **Descriptor Track**: generate useful vector representations of videos for this video similarity task.
- **Matching Track**: create a model that directly detects which specific clips of a query video correspond to which specific clips in one or more videos in a large corpus of reference videos.

https://www.drivendata.org/competitions/group/meta-video-similarity/

A turnkey solution: Hasher-Matcher-Actioner

Hasher-Matcher-Actioner (HMA) is an

- Open-source (github.com/facebook/ThreatExchange/tree/main/hasher-matcher-actioner),
- turnkey,
- trust and safety tool.

A turnkey solution: Hasher-Matcher-Actioner

Hasher-Matcher-Actioner (HMA) is an

- Open-source (github.com/facebook/ThreatExchange/tree/main/hasher-matcher-actioner),
- turnkey,
- trust and safety tool.

You can

- submit content to your own instance of HMA to scan through content on your platform
- flag potential community standards violations.
- configure rules in HMA to automatically take actions (such as enqueue to a review system) when these potential violations are flagged.

A turnkey solution: Hasher-Matcher-Actioner

Hasher-Matcher-Actioner (HMA) is an

- Open-source (github.com/facebook/ThreatExchange/tree/main/hasher-matcher-actioner),
- turnkey,
- trust and safety tool.

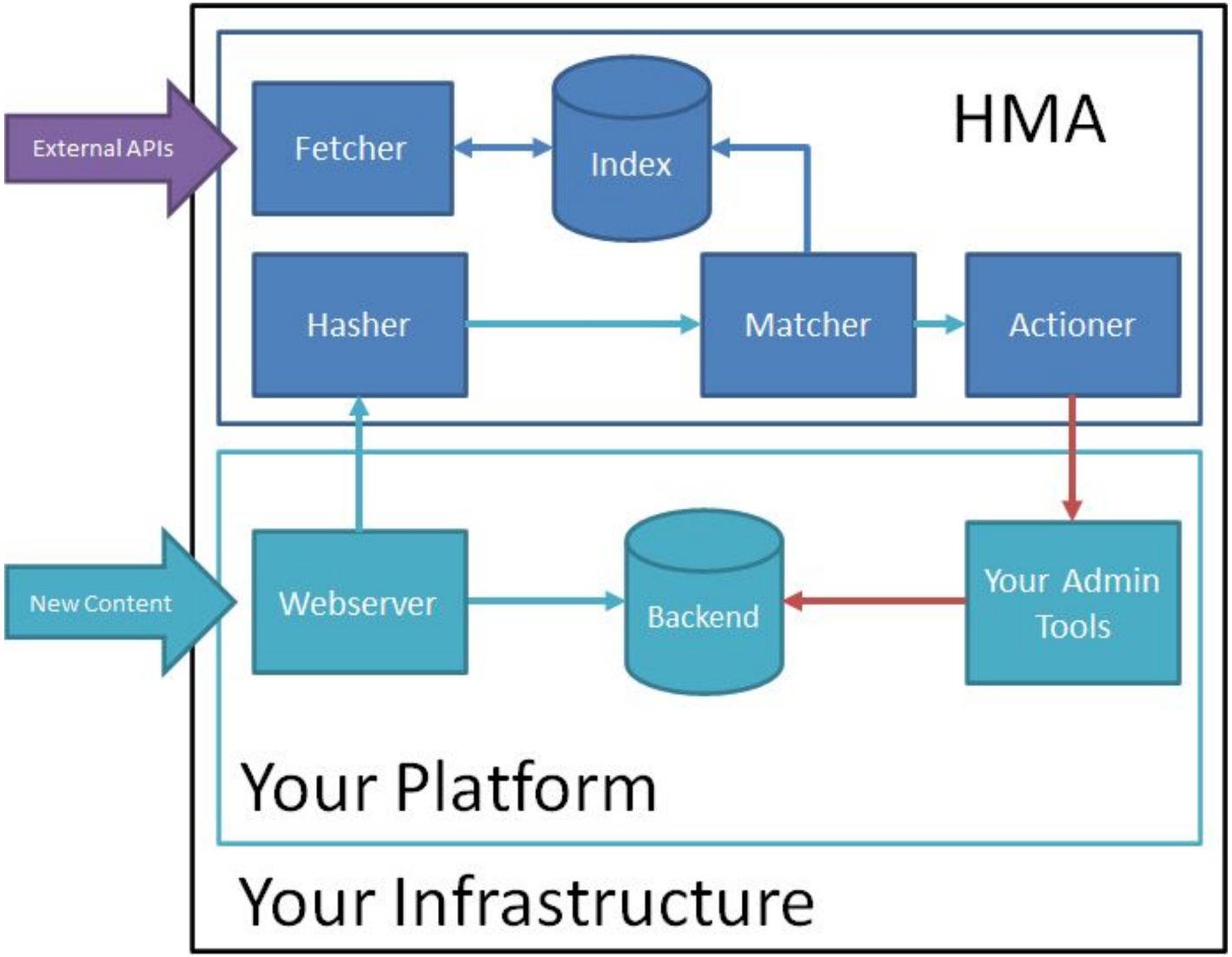
You can

- submit content to your own instance of HMA to scan through content on your platform
- flag potential community standards violations.
- configure rules in HMA to automatically take actions (such as enqueue to a review system) when these potential violations are flagged.

HMA

- can pull in violating seeds from Facebook's ThreatExchange API (upload yours too!).
- works on AWS only (heavily uses Lambda to minimize cost), Terraform available.

HMA Architecture



Wrapping up

- Automation is necessary to be effective, but you will lose precision. Human support always needed for appeals and ground truth. Do expect false positives.
- **PDQ**, **VideoPDQ**, **VideoMD5** and **SSCD** provide you with a way to obtain compact representations.
- **HMA** provides you with a turnkey solution to search those representations and enforce Integrity.
- **ThreatExchange** provides you with a platform for exchanging representations.


Meta