MAAM: Media Asset Annotation and Management

Manos Schinas
Centre of Research and Technology - Hellas
Thessaloniki, Greece
manosetro@iti.gr

Panagiotis Galopoulos
Centre of Research and Technology - Hellas
Thessaloniki, Greece
gpan@iti.gr

Symeon Papadopoulos
Centre of Research and Technology - Hellas
Thessaloniki, Greece
papadop@iti.gr

ABSTRACT

Artificial intelligence can facilitate the management of large amounts of media content and enable media organisations to extract valuable insights from their data. Although AI for media understanding has made rapid progress over the recent years, its deployment in applications and professional sectors poses challenges, especially to organizations with no AI expertise. This motivated the creation of the Media Asset Annotation and Management platform (MAAM) that employs state-of-the-art deep learning models to annotate and facilitate the management of image and video assets. Annotation models provided by MAAM include automatic captioning, object detection, action recognition and moderation models, such as NSFW and disturbing content classifiers. By annotating media assets with these models, MAAM can support easy navigation, filtering and retrieval of media assets. In addition, our platform leverages the power of deep learning to support advanced visual and multi-modal retrieval capabilities. That allows accurately identifying assets that convey a similar idea, or concept even if they are not visually identical, and support a state-of-the-art reverse search facility for images and videos.

ACM Reference Format:
Manos Schinas, Panagiotis Galopoulos, and Symeon Papadopoulos. 2023. MAAM: Media Asset Annotation and Management. In *ICMR '23: International Conference on Multimedia Retrieval Proceedings, June 12–15, 2023, Thessaloniki, Greece.* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The need for digital asset management tools arises due to the increasing amount of media content produced by media organizations. As the volume of media assets grows, it becomes more challenging to effectively manage and organize them. Digital asset management tools help streamline the process of analyzing, storing, organizing and finally retrieving assets, making it easier for organizations and users to maximize the value of their content. While existing solutions provide basic functionalities for organizing and retrieving assets, the cost of using them can be a significant burden. It is worth noting that the majority of DAM solutions come with a high cost due to commercial licensing. Additionally, they may not always meet the needs of an organization, when it comes to categorizing and annotating assets. The ability to annotate and categorize assets helps improve retrieval and enables organizations to extract insights from their media content. However, the majority of organizations and solutions rely on commercial computer vision APIs that increase even further the cost, especially for smaller organizations. To this end, the Media Asset Annotation and Management platform (MAAM) is a powerful state-of-the-art solution for media organizations to manage, annotate, and extract insights from their vast amounts of image and video assets. MAAM can be deployed on one's own infrastructure, offering greater security and privacy, since the organization has full control over the storage and management of its data. Also, with the use of advanced deep learning models, MAAM allows organizations to add valuable metadata to their assets, making it easier for them to filter and retrieve their content. One of the key features of MAAM is its ability to annotate media assets with various AI models, including captioning, object detection, action recognition, and content moderation models. This rich set of annotations makes it easier for users to categorize, understand, and manage their media assets. Moreover, MAAM leverages the power of deep learning to understand the visual characteristics and context of media assets, making it possible to identify assets that convey a similar message, represent a similar idea, or illustrate a similar concept even if they are not visually identical. In addition, MAAM also offers a cutting-edge reverse search facility for images and videos. A working prototype can be found online in https://maam.mever.gr/.

2 OVERVIEW OF MAAM PLATFORM

MAAM provides advanced asset management, organization, and retrieval capabilities that can be used in several usage scenarios. The platform supports the upload of image and video media assets and has the capability to annotate them in near real-time. The annotations can be leveraged as filters when searching for assets. In addition to annotations, the platform also supports free-text search for querying and discovering assets, based either on metadata provided by the user (e.g. title, description and filename) or on metadata generated by MAAM with the help of AI models (e.g. image captions). This combination of annotation-based filtering and free-text search enables users to quickly and easily find the media assets they need by using relevant keywords or categories. For the same reason, MAAM offers advanced visual similarity features that allow users to retrieve content based either on semantic concepts or on strict visual-based search. The semantic visual similarity search is based on the dense vector representation extracted using the CLIP model [8], which encodes the semantic information of the visual content of the asset. By using approximate k-nearest neighbor search, users can retrieve semantically similar content

ICMR '23, June 12–15, 2023, Thessaloniki, Greece

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

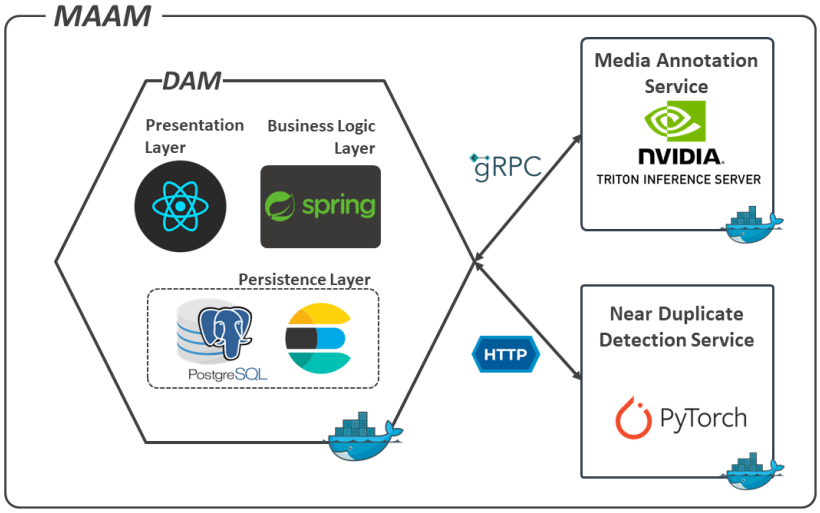

Figure 1: MAAM Service-oriented Architecture

and discover new assets of interest. Moreover, MAAM offers a reverse image search functionality, which can be used to identify near-duplicate content, such as assets that are visual variations of each other. This allows organizations to identify media assets that are similar to one another, even if they are not exact duplicates. This includes visual variations of images as well as partially matching videos, making it easier for organizations to identify duplicates of their content, detecting copyright violations, and much more. With the combination of advanced visual similarity features and annotations, users can easily retrieve and discover media assets based on their specific needs and requirements, making MAAM a powerful tool for managing and organizing media content. By using these advanced retrieval features, the annotated assets can be organized into collections, referred to as "projects" in MAAM, and these projects allow users to effectively categorize their content.

2.1 Platform Architecture

MAAM is a service-oriented solution, consisting of three loosely interrelated services: the Digital Asset Management service (DAM), the Media Annotation Service, and the Near-Duplicate Detection (NDD) service. The DAM service is the centerpiece of the MAAM platform, as it offers a robust set of core functionalities, ranging from user authentication and authorization to asset organization. Its presentation layer consists of a React-based user interface (UI), that provides a user-friendly and intuitive interface for accessing MAAM's main functionalities. The platform's business logic is implemented by a Spring-based application that follows the Model-View-Controller (MVC) architecture pattern. The persistence layer of MAAM is responsible for storing and maintaining the digital assets, their associated metadata and any other entity needed in the application. This layer relies on two storage frameworks, each serving a specific purpose: PostgreSQL and Elasticsearch. PostgreSQL, an efficient relational database management system, is used as the primary application storage. Elasticsearch, on the other hand, serves as a search and analytics engine that enables efficient searching of digital assets. The DAM leverages Elasticsearch's ability to perform full-text search, allowing users to find assets based on keywords and phrases contained in the metadata. Also, the aggregation feature of Elasticsearch is used for faceted search based on specific attributes such as asset type, upload time, annotation type, etc. In addition, Elasticsearch supports more advanced and sophisticated search functionalities, such as k-nearest neighbor (kNN) search based on dense feature vectors generated by the Media Annotation Service. The Media Annotation Service is responsible for hosting and managing the AI annotation models described in the next section, and it is a key component of the platform as it provides unique features that differentiate MAAM from other asset management tools. The Media Annotation Service is based on the NVIDIA Triton Inference Server[1], which provides a flexible and scalable solution for deploying AI models in a production environment. Triton can host AI models implemented in most of the major frameworks, such as TensorFlow, PyTorch, and ONNX, making MAAM easily extendable with new state-of-the art models. By leveraging Triton, the Media Annotation Service can run AI models very efficiently, making it possible to perform almost real-time annotation of media assets as these are uploaded to the platform. The communication between the DAM service and the Media Annotation Service is facilitated by gRPC [2], a high-performance framework for remote procedure calls. This asynchronous communication is particularly important for MAAM, as the Media Annotation Service hosts multiple AI models with varying levels of complexity and processing requirements. The use of gRPC enables the DAM service to send multiple assets to the Media Annotation Service for annotation, without waiting for the annotation process to complete. The Media Annotation Service can perform the annotation in parallel, without blocking the overall asset management process and affecting its performance. As

[1]https://developer.nvidia.com/nvidia-triton-inference-server
[2]https://grpc.io/

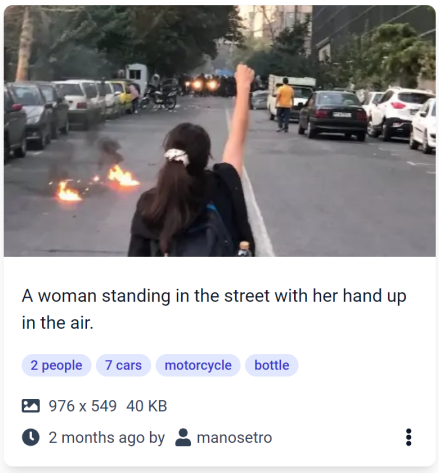

Figure 2: An example of an asset card, for an image uploaded in MAAM with the corresponding annotations

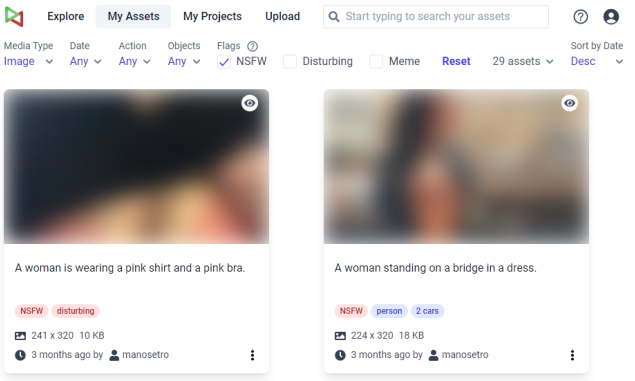

Figure 3: Assets page with NSFW tag enabled and media type filter being image

each asset is annotated by one of the supported models, the results are sent back to the DAM and the asset's metadata are updated asynchronously.

3 ANNOTATION MODELS

MAAM provides a set of annotation models that are automatically applied to all the uploaded images and videos. The resulting annotations are stored in PostgreSQL and indexed in Elasticsearch, making them easily accessible and searchable. This allows MAAM to offer a wide range of annotations, which can also be customized according to specific requirements. If new annotation models are needed, they can be easily deployed in the Media Annotation Service as described in 2.1 and a new handler can be added in the DAM side to process and store the new results. This makes it simple to extend the capabilities of MAAM to meet changing needs and requirements. In the current version of MAAM we have included a set of state-of-the-art models ranging from image captioning to object, face and action detection. Given the large number of models integrated in MAAM, we used the InDistill model compression approach [10] to reduce the size and inference time of several of these models. InDistill combines knowledge distillation and channel pruning in a unified framework for the transfer of the critical information flow paths from heavyweight teachers to lightweight student models.

Automatic image captioning: The captioning model generates descriptive text for each image. We use OFA [14], a state-of-the-art captioning model, and the text produced by it is then indexed in Elasticsearch, enhancing the findability of visual content. An example of a generated caption is depicted in Figure 2.

Object detection: The object detection model is used to identify objects within images and video frames. We use Faster R-CNN [9] with an InceptionV2 [4] backbone, trained on the 80 object classes of the MS COCO dataset [7]. In case of images, we detect and store the bounding box containing the corresponding object, while in videos we also provide time information. In both cases, a confidence score is also included. If a user wants to find images containing a specific object, they can filter assets by using the objects filter and the platform will return all images that have been annotated with that label. Figure 2 provides an example of objects in an image. he model can detect even small objects that cover a small part of the image, such as the small plastic bottle in the woman's backpack.

Action recognition: For action recognition in videos, we consider the SlowFast R50 model [2] trained on the Kinetics400 dataset [12] [3]. For images, due to the fact that Kinetics400 contains videos and most action recognition models use 3D CNNs, we used a ResNet152 model [3] with the TSN approach [13] for training at frame level. At inference, we directly apply the frame level classifier to the images.

Face Detection and Recognition: For face detection we consider the VGGFace2 model [1] trained on about 9k faces of the VGGFace2 dataset [4]. In case of videos, we apply the same model on randomly selected video keyframes.

Content Moderation: The MAAM platform employs two moderation models to filter content, ensuring the platform remains safe and appropriate for all users. We trained two moderation models using an iterative approach that leverages large image datasets in a semi-automated annotation scheme [11]. The resulting models are able to detect disturbing and Not Safe For Work (NSFW) content. For images, the moderation results, including a confidence score, are stored and indexed. For videos, the models are applied to keyframes and a video is considered NSFW or disturbing if at least one scene receives such a tag. At retrieval time, users have the option to include or exclude that type of content, while the UI uses these tags to blur the corresponding asset. Users can then choose to reveal the actual content at their own discretion. For example, in Figure 3, the user has enabled the NSFW filter to get only those that have been tagged as such.

Meme Detection: For images, we determine whether they are memes or not using our previous MemeTector model [6]. We apply the model to each image and the result is stored and indexed along with a confidence score.

[3] https://github.com/facebookresearch/SlowFast
[4] https://paperswithcode.com/dataset/vggface2-1

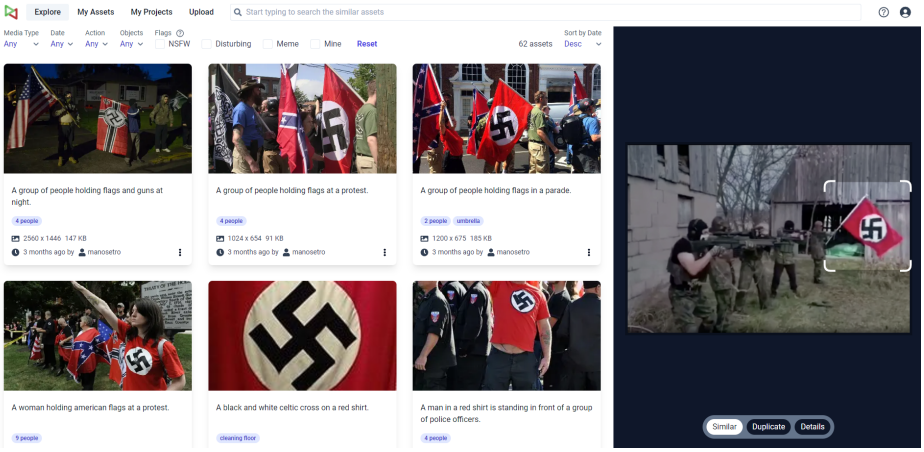

Figure 4: Find similar asset functionality by selecting a region in an image

Visual Similarity: To support image similarity, we utilized the pre-trained CLIP model [8], which was trained on a large image-text dataset, consisting of around 400 million pairs. We used the ViT-B/32 version of the image encoder[5]. For each image uploaded in the DAM, MAAM receives from the Media Annotation Service a dense vector representation of 512 dimensions that encodes the semantic information of the image's visual content. This is then indexed in Elasticsearch and used to retrieve semantically similar content through the approximate k-nearest neighbor (kNN) search feature of Elasticsearch. This feature can be used in conjunction with all other features of Elasticsearch, allowing kNN queries to be combined with free text searches, filters, and aggregations. This flexibility enhances the user experience by allowing users to refine further the visual similarity search results to find content of specific interest. In addition to global image-level similarity, we support region-based retrieval. This allows users to select a specific region in an image by defining a bounding box, and retrieve content that is visually similar to that region in a more focused manner. To achieve this, the platform extracts a dense vector representation of the selected region in real-time, as it does for entire images during upload. This representation is then used in the same kNN search process to find relevant content. An example of the visual similarity feature is depicted in Figure 4: by selecting an image region containing a swastika, the platform retrieves assets containing the same symbol, as The dense vector extracted with CLIP encodes the meaning of the swastika symbol.

4 REVERSE IMAGE AND VIDEO SEARCH

The Near-Duplicate Detection (NDD) service provides a highly efficient reverse search functionality, helping users to quickly identify and remove duplicate assets within their collections. NDD is based on the Distill-and-Select framework (DnS) [5] in order to provide efficient and accurate indexing and retrieval of images and videos. The former functionality analyses the provided multimedia item based on their visual content and adds them to the corresponding index. The latter functionality searches the constructed index for near-duplicates to a query multimedia item and ranks the retrieved results based on their similarity to the query. The service provides calls to: (i) add images and videos to the corresponding indexes by providing their explicit URLs, (ii) search the index for near-duplicates given a query multimedia item, providing several options for similarity calculation, and (iii) create collections of multimedia items for the better organization and search of near-duplicates. Internally, the NDD follows a service-oriented architecture, consisting of several modular services for feature extraction, indexing, and searching, while the communication between the NDD service and other components of the MAAM platform is performed through a REST API, exposed by the NDD service.

5 FUTURE WORK

MAAM is under development towards enhancing its features and performance. In the future, the development team aims to incorporate more models such as OCR, improve the performance of the existing models or fine-tune their predictions. For example, in case of moderation models, we are investigating ways to support more fine-grained classes, compared to the current broad categories like NSFW and disturbing. Furthermore, we plan to include few-shot learning capabilities, enabling users to define their own classes and objects, by selecting only a few examples from their assets to act as a support set. Then these new classes be used as filters in asset retrieval. Moreover, we are working towards adding user management features such as creating groups of users with different access permissions, enabling multiple users to collaborate in organizing assets. We also plan to add user feedback features, allowing users to provide new annotations, correct existing ones, or provide other types of feedback. This mechanism can be useful in using MAAM as a labeling tool for facilitating the development or fine-tuning of AI models.

ACKNOWLEDGMENTS

This work is partially funded by the Horizon 2020 European project MediaVerse under Grant Agreement no. 957252.

[5]https://github.com/openai/CLIP

REFERENCES

[1] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. 2018. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 67–74.

[2] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6202–6211.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[4] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*. pmlr, 448–456.

[5] Giorgos Kordopatis-Zilos, Christos Tzelepis, Symeon Papadopoulos, Ioannis Kompatsiaris, and Ioannis Patras. 2022. DnS: Distill-and-Select for Efficient and Accurate Video Indexing and Retrieval. *International Journal of Computer Vision* 130, 10 (2022), 2385–2407.

[6] Christos Koutlis, Manos Schinas, and Symeon Papadopoulos. 2022. MemeTector: Enforcing deep focus for meme detection. *arXiv preprint arXiv:2205.13268* (2022).

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 740–755.

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015).

[10] Ioannis Sarridis, Christos Koutlis, Symeon Papadopoulos, and Ioannis Kompatsiaris. 2022. InDistill: Transferring Knowledge From Pruned Intermediate Layers. *arXiv preprint arXiv:2205.10003* (2022).

[11] Ioannis Sarridis, Christos Koutlis, Olga Papadopoulou, and Symeon Papadopoulos. 2022. Leveraging Large-scale Multimedia Datasets to Refine Content Moderation Models. In *2022 IEEE Eighth International Conference on Multimedia Big Data (BigMM)*. IEEE, 125–132.

[12] Lucas Smaira, João Carreira, Eric Noland, Ellen Clancy, Amy Wu, and Andrew Zisserman. 2020. A short note on the kinetics-700-2020 human action dataset. *arXiv preprint arXiv:2010.10864* (2020).

[13] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*. Springer, 20–36.

[14] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*. PMLR, 23318–23340.