Natural Language Database Querying using Image Recognition and Natural Language Processing

Aditeya Baral
BTech CSE
PES University
aditeya.baral@gmail.com

Vishesh P
BTech CSE
PES University
visheshp172000@gmail.com

Anirudh Haritas Murali
BTech CSE
PES University
anihm136@gmail.com

Vinay V Kirpalani
BTech CSE
PES University
vinaykirpalani@gmail.com

***Abstract*—We present a novel approach to data retrieval from tagged databases using only natural language audio queries. The approach is capable of handling multiple languages and accents and retrieves data in linear time. We have used the concept of semantic vector spaces and word embeddings to perform rapid retrieval and matching of queries at a semantic level. We also present a holistic and semi-supervised approach to generate tags for any given database of videos or images. It combines both image recognition and natural language processing to identify objects and spoken entities to generate a set of words that identify the theme and content of the video.**

***Index Terms*—database, image recognition, natural language processing, audio query**

I. BACKGROUND

With the voluminous amount of data generated everyday that is growing at an exponential rate, easy and efficient retrieval of relevant data becomes increasingly important. The growth of video platforms and their catalog of diverse videos has made it exceedingly difficult to manually tag each of these videos for retrieval. The current need of the hour for such platforms is a system that is capable of retrieving relevant matching videos for a given query efficiently and is easily scalable.

Current state of the art approaches include like Visual Semantic Reasoning [1], Central Similarity Quantization [2], Multiple Visual-Semantic Embeddings [3] and Joint Text-Video Embeddings [4]. These approaches use deep learning architectures to match a video to an audio or text query directly. The disadvantage of these approaches is that being supervised approaches, they rely heavily on the quality and diversity of the training data to generalise well for multiple types of videos, voices and languages. Training such models on a large dataset would require a large amount of time and computing resources. The proposed idea uses unsupervised and semi-supervised approaches instead eliminating the need for any training dataset and also speeds up the overall execution and retrieval process.

II. APPROACH

The proposed approach combines image recognition along with natural language processing to retrieve videos matching a natural language query from a video database. The retrieval process consists of three major steps - recognition of the natural language audio query, extraction of features and tags from the database of videos and matching of the query to a video.

The input audio query is first transcribed and then translated into English. This allows us to tackle multiple natural languages and accents with the same approach. Google's Speech Recognition API [5] is deployed to perform rapid transcription and translation of the audio query. IBM Watson API is then used to extract features from the audio query such as the keywords of the transcript as well as the entities mentioned. An audio processing pipeline is created to simultaneously perform the translation as well as the extraction of features.

The database of videos is pre-tagged before the audio query is processed. Each video is split into frames, which are then filtered based on image features [6] to retain frames depicting every scene in the video. A Google reverse image search is performed on every filtered frame, to lookup the same video on any online platforms such as YouTube and extract tags from these platforms. The frame is then analysed to look for objects and entities [7], which are extracted and stored. Further, a language model is used to generate a textual description of each frame. This generated description is searched on Google and suggested relevant searches are retrieved. Additionally, keywords from this description are extracted and added to the full set of results obtained.

All the obtained image based tags are then filtered to retain only the top frequently occurring tags across all frames. These filtered tags are combined with the keywords and entities in the audio of the video to form the final set of tags for a video. A vector space is created out of these tags for every video in the dataset [8]. The vector feature spaces obtained help represent the video's theme at a semantic level. Similarly, the query is also transformed into a vector space by using the keywords obtained earlier. Finally, the query's feature space is compared with the feature spaces of all the videos in linear time, using cosine similarity as the metric of evaluation. The video with the highest similarity is retrieved and displayed to the user.

III. PERFORMANCE

The proposed approach displays high performance with extremely low latencies. The performance was measured against the top 50 videos of the MSR-VTT [9] dataset. Each video in the dataset has multiple descriptions and has an overall runtime of under 15 seconds. Multiple videos were considered with

different queries describing each video and the performance of our model was measured using these queries as natural language inputs. Our model was able to retrieve matching videos with an average latency of 0.0203 seconds per query.

The model when deployed against the dataset is able to record an accuracy of 20% for precise video matches and an accuracy of 32% for same category matches. The low accuracies on the benchmark dataset may be attributed to the very short lengths of the videos which affect the semi-supervised based approach.

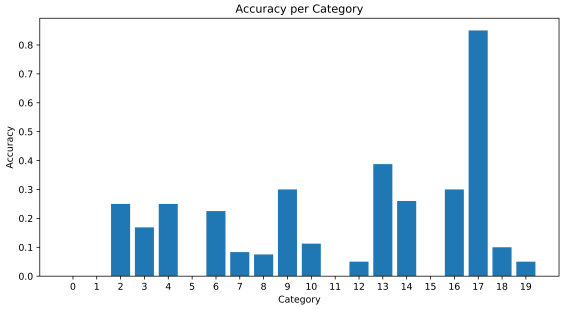

Fig. 1. Accuracy vs Category

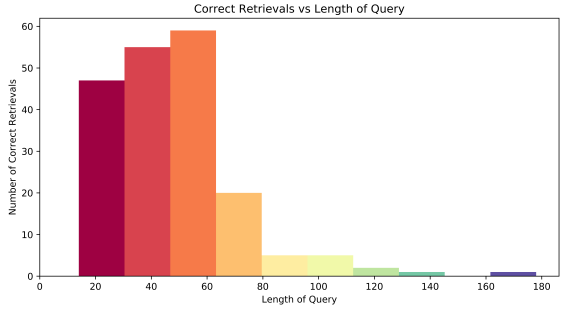

Fig. 2. Correct Retrievals vs Length of Query

REFERENCES

[1] Z. Feng, Z. Zeng, C. Guo, and Z. Li, "Exploiting visual semantic reasoning for video-text retrieval," 2006. [Online]. Available: https://arxiv.org/pdf/2006.08889v1.pdf

[2] L. Yuan, T. Wang, X. Zhang, F. E. Tay, Z. Jie, W. Liu, and J. Feng, "Central similarity quantization for efficient image and video retrieval," 2019. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/papers/Yuan_Central_Similarity_Quantization_for_Efficient_Image_and_Video_Retrieval_CVPR_2020_paper.pdf

[3] H. M. Nguyen, T. Miyazaki, Y. Sugaya, and S. Omachi, "Multiple visual-semantic embedding for video retrieval from query sentence," 2004. [Online]. Available: https://arxiv.org/pdf/2004.07967v1.pdf

[4] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivi, "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips," 2019. [Online]. Available: https://arxiv.org/pdf/1906.03327v2.pdf

[5] A. Zhang. (2017) Speech recognition (version 3.8). [Online]. Available: https://github.com/Uberi/speech_recognition

[6] J. Buchner, "Imagehash," https://github.com/JohannesBuchner/imagehash, 2020.

Fig. 3. Accuracy 100

[7] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[8] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50, http://is.muni.cz/publication/884893/en.

[9] T. Y. Jun Xu, Tao Mei and Y. Rui, "MSR-VTT," https://www.microsoft.com/en-us/research/publication/msr-vtt-a-large-video-description-dataset-for-bridging-video-and-language/, 2016.