Semantic Video Retrieval using Deep Learning Techniques

Danish Yasin
Bahria University
Islamabad, Pakistan
danishyasin33@gmail.com

Ashbal Sohail
Bahria University
Islamabad, Pakistan
ashbalsohail@gmail.com

Imran Siddiqi
Bahria University
Islamabad, Pakistan
imran.siddiqi@bahria.edu.pk

***Abstract*—Content based video retrieval has been an active research area for many decades. Unlike tagged-based search engines which rely on user-assigned annotations to retrieve the desired content, content based retrieval systems match the actual content of video with the provided query to fetch the required set of videos. Thanks to the recent advancements in deep learning, the traditional pipeline of content based systems (pre-processing, segmentation, object classification, action recognition etc.) is being replaced by end-to-end trainable systems which are not only effective and robust but also avoid the complex processing in the conventional image based techniques. The present study exploits these developments to develop a semantic video retrieval system accepting natural language queries and retrieving the relevant videos. We focus on key individuals appearing in certain scenarios as queries in the current study. Persons appearing in a video are recognized by tuning FaceNet to our set of images while caption generation is exploited to make sense of the scenario within a given video frame. The outputs of the two modules are combined to generate a description of the frame. During the retrieval phase, natural language queries are provided to the system and the concept of word embeddings is employed to find similar words to those appearing in the query text. For a given query, all videos where the queried individuals and scenarios have appeared are returned by the system. The preliminary experimental study on a collection of 50 videos reported promising retrieval results.**

***Index Terms*—Semantic Retrieval, Deep Convolutional Neural Networks (CNNs), Long-Short Term Memory Networks (LSTMs)**

I. Introduction

The last few years have witnessed a tremendous increase in the number of imaging sensors thanks to the recent advancements in the imaging technology. Consequently, the amount of digital multimedia data in the form of images, audios and videos has increased manifolds. Video archives like YouTube, DailyMotion and Vimeo etc. receive hundreds of hours of videos every minute and billions of users every single day. Such huge collections must be complemented with smart retrieval solutions allowing users efficiently search the desired content. Traditionally, most of the archives employ tag-based search schemes. Such solutions are trust-based systems where users are entrusted to upload accurate and relevant description of their content. Even in cases where tags are correctly assigned, a small set of descriptive attributes cannot effectively capture the rich information in the visual content. This was the key motivation that guided the research endeavors in content based retrieval solutions which exploit the actual content within the videos for retrieval purposes.

Supported by IGNITE, National Technology Fund, Pakistan.

Content based video retrieval (CBVR) systems have gained significant research attention in the last two decades. Many of the problems in this domain have been revisited thanks to the recent advancements in deep learning and availability of large labeled datasets. Content in videos can be categorized into three broad categories, the visual content (objects, persons, buildings etc.), the textual content (News tickers, score cards etc.) and the audio content. Among these, visual content represents the richest form of information and is the most common choice for retrieval queries. From the view point of level of abstraction and complexity of system, three levels of content based retrieval systems have been identified in the literature.

- **Level-I:** Retrieval of content using basic visual features such as color, texture and shape. An example query of what a Level-I based system can accept could be to retrieve *'Videos with red circles'*.
- **Level-II:** Retrieval of content through detection of objects adding a degree of 'smartness' in the system. An example query for such systems could be to find *'Videos of cars'*.
- **Level-III:** Retrieval of content through abstract and high-level queries is the focus of Level-III systems. Examples could be to search videos having *'A group of people protesting'* or videos where *'Spider man saves the lady'*.

In general, the Level-II of content based retrieval framework is known as *'conceptual retrieval'*, whereas Level-III is commonly termed as *'semantic retrieval'*. Conceptual retrieval has limited success when dealing with queries that may involve complex high-level concepts and this gap between the two levels is known as the semantic gap. The recent advancements in machine learning have allowed researchers to bridge this gap replacing the conventional segmentation and recognition techniques by end-to-end learning systems. Our present work is also an attempt in this direction. We aim to develop a robust semantic (Level-III) video retrieval system that is able to accept high level natural language queries from the users and retrieve relevant

978-1-7281-4676-8/20/$31.00©2020 IEEE

videos. More specifically, among various categories of visual content we are interested in 'persons' and their corresponding 'actions'. The proposed system employs state-of-the-art deep learning architectures to recognize individuals as well as actions within videos. Dedicated models are trained for each of the two tasks i.e. a separate model to recognize faces and another one to identify the possible actions. The outcomes of the two models are then combined to assign a meaningful description to the processed frames in a video. During the retrieval phase, a user may provide a natural language query like *'Donald Trump meeting Vladimir Putin'* and the system is expected to retrieve the relevant videos.

The paper is organized as follows. We first present an overview of the notable contributions to content based video retrieval in Section II. Section III presents the details of the proposed technique introducing the developed models. Experimental results and an analysis of the key findings are summarized in Section IV while Section V concludes the paper.

II. Related Work

Visual content based retrieval of images and videos has been an attractive research area for many years. A number of indexing and retrieval systems have been proposed for image as well as video databases. Among well-known retrieval systems, QBIC [4] developed by IBM in the mid 1990s is known to be the first commercial Content Based Image Retrieval (CBIR) system. The system supports image retrieval using color, shape or texture and the user may provide a single or multiple features in the query. Other classical systems include Photobook [14], VisualSEEK/WebSEEK [21], I-Motion [19] and I-Match [22] etc. Content based intelligent retrieval of videos (CBVR) offers even a wider spectrum of applications as compared to CBIR. The importance of CBVR is reflected by the fact that since 2003 the National Institute of Standards and Technology (NIST) and a number of government agencies in the United States have been regularly sponsoring the TREC (Text Retrieval Conference) Video Retrieval Evaluation (TRECVid) [16]. TRECVid provides a large collection of videos and every year a number of video retrieval algorithms are submitted for evaluation and comparison.

A comprehensive information retrieval system is presented in [1] allowing recognition of semantic concepts in videos. The technique relies on adaptation of evident theory to neural networks along with an exploitation of the relationship between different concepts and descriptors and, ontology based concepts. Authors term the system as Ontological-PENN and evaluate it on the TRECVid dataset to demonstrate the effectiveness of the proposed retrieval scheme. In another notable work, a content based video retrieval system to search relevant videos in the German Broadcasting Archive is presented in [12]. The system mainly relies on deep neural networks and in addition to similarity search, offers scene change detection, video OCR and person detection features.

The advent of deep learning has bridged the gap between computer vision and machine learning. Conventional object detectors and descriptors (for recognition) have been replaced by machine learned object detectors [18], [7], [17] and data driven feature extractors [9], [3]. In the context of content based search, a number of recent studies validate the effectiveness of deep learning based techniques for smart image [27], [24] as well as video retrieval [15], [28], [8]. Tzelepi et al. [24], for instance, employ a deep CNN for feature extraction to support content based image retrieval. The extracted features are converted to compact image descriptors to reduce the memory requirements. Authors in [10] employ a deep learning framework to learn binary codes for the videos taking into account the temporal as well as discriminative information within a video resulting in an effective hashing scheme. Jabeen et al. [8] extracted a set of descriptors using RNNs to categorize videos into various genres including music, sports, meeting etc. Video summarization using deep semantic features is considered in [13] where features extracted by deep CNNs are clustered in the semantic space to identify similar segments in videos. In another notable work, Venugopalan et al. [25] map videos to natural language sentences using convolutional as well as recurrent neural networks. In a series of related studies [29], [6], [5], authors employ CNNs with attention based LSTMs for video captioning. Once the captions are generated, they can be exploited for semantic retrieval using natural language queries.

It can be noticed from the above discussion that deep learning based solutions have dominated the research in content based image and video retrieval during the last few years. Not only such systems are robust, they avoid many of the steps in a classical pipeline (pre-processing, segmentation and recognition) allowing training of end-to-end systems. The studies mentioned above pave way for a system such as ours to be developed. The content based video retrieval system of the German broadcasting archive had been an inspiration for this study. However, the German system lacked the ability to retrieve videos using semantic search. Our study brought smartness to such content retrieval systems through the use of natural language processing and robust search algorithms. Our study is inspired by the robustness of such deep networks that can be effectively exploited for smart indexing and retrieval of videos as detailed in the following section.

III. Methods

This section presents the technical details of the proposed retrieval system, an overview being presented in Figure 1. The system mainly comprises two key components, person recognition and caption generation (action recognition). Since successive frames in a video contain redundancies, we extract one frame every two seconds for processing. Each frame is fed to the two modules of the system. The person recognition

module employs the state-of-the-art FaceNet [20] by fine tuning it to the dataset under study. For action recognition, we exploit the caption generation framework [29]. The outputs of both the modules are then combined and, identified persons and corresponding actions (if any) are stored in a database. The database serves as an index (look-up table) containing frame locations corresponding to the persons of interest as well as metadata on the videos in which they have appeared. When a natural language query is presented to the system, it parses the query string, identifies names of persons and actions and retrieves the relevant videos using word embeddings. Details of these steps are presented in the following.

A. Person Recognition

As discussed earlier, identification of key persons in videos relies on face recognition for which we tune the pre-trained FaceNet implementation [20]. Previous face recognition implementations using deep learning were highly inefficient in comparison to FaceNet, which directly trains its output to be a compact 128 dimension embedding by using a triplet loss function. The functions consists of two matching and a non-matching face thumbnails, where the function aims to separate the positive from the negative by a euclidean distance measure. We fine tuned this distance margin to perform optimally for our system. The model employs Inception ResNet v1 architecture [23] and is trained on the VGGFace2 database. The database has more than 3.3 Million face images belonging to around 9000 different individuals with more than 360 images per subject. We employ the pre-trained model in a transfer learning framework and tune it to the facial images in our dataset. For each individual we considered 300 training images. More details on the dataset are presented in Section IV of the paper.

B. Action Recognition

Recognition of actions and generation of captions from videos and images has been investigated for a long time. Contrary to the complex image analysis based techniques, the deep learning based pipeline exploiting both convolutional and recurrent neural networks has proven to be very effective in a number of studies [25], [29], [6], [5]. We have employed the same idea to generate captions from individual video frames using the framework presented in [29] trained on the MS COCO (Common Objects in Context) dataset [2], [26]. The framework comprises of three distinct steps. A ResNet-101 encoder comes first that has been trained using transfer learning, the encoder can be fine tuned to improve its performance. Then comes the LSTM based decoder, which is responsible for generating the captions. Lastly, we have the attention network that calculates the weights of the image and determine which part of the image is most relevant for the creation of a caption. Once the caption is generated, the stop words are removed and the words (corresponding to possible scenarios) are kept for further processing.

Figure 2 illustrates an example where a video frame is fed to the system (both the modules). The person recognition module identifies the individuals and outputs *'Donald Trump'* and *'Emmanuel Macron'*. Likewise, the caption generator outputs the caption: *'Two men standing next to each other'*. Corresponding to the frame, an entry is made in the database with the names of the two individuals and the verb *'standing'*. The process is repeated for all the frames (we process one frame every two seconds) and the database is updated accordingly with the meta-data of the video, time stamps of frames and the identified individuals and actions.

C. Retrieval

During the retrieval phase, user provides a natural language query which is parsed to identify individuals and scenarios. While the individual names are likely to be unique, the user may supply any of the words in the high level query which may not necessarily be the same as those stored in the database. We, therefore, employ word embeddings that are produced using the Word2Vec [11] model. Word embeddings are vector representations of words in such a way that similar words are likely to appear in close proximity of one another in the projected vector space. In other words, the cosine similarity between the vectors representing similar words is expected to be high. We employ the same idea to obtain k closest matches to the provided words in the query ($k = 5$ in our experiments). This allows more flexibility in the way users formulate queries. As an example, for the query *'meeting'*, the five nearest words (with respect to similarity score) come out to be *'Conference: 0.909'*, *'Meetings: 0.882'*, *'Talks: 0.867'*, *'Discuss: 0.863'* and *'Meet: 0.861'*.
Significant effort was put into developing a search algorithm for effective retrieval. Firstly, we designed the algorithm while keeping in mind the probability of false positives in the recognition of faces. Moreover, by using Word2Vec, we managed to improve the retrieval of accurate videos by up to 20% as direct word matching was rare and many videos were missed. To counter false positives, the search algorithm considered the window of five frame entries around the detection of the person being searched. If the relative probability of that person in the eleven frames is the highest then it is considered a true positive. A check was also put into place that avoided recursive checking of frames, so that if a detection is found to be true positive then the next five frames were skipped for the detection of the same person.

IV. Results and Analysis

For experimental study, we collected a set of 50 videos from YouTube. The videos were chosen so as to contain key individuals considered in our dataset. Being a pilot study, we initially focused on ten different key personalities as summarized in Figure 3. For tuning the FaceNet, we collected 300 images of each of these individuals. These

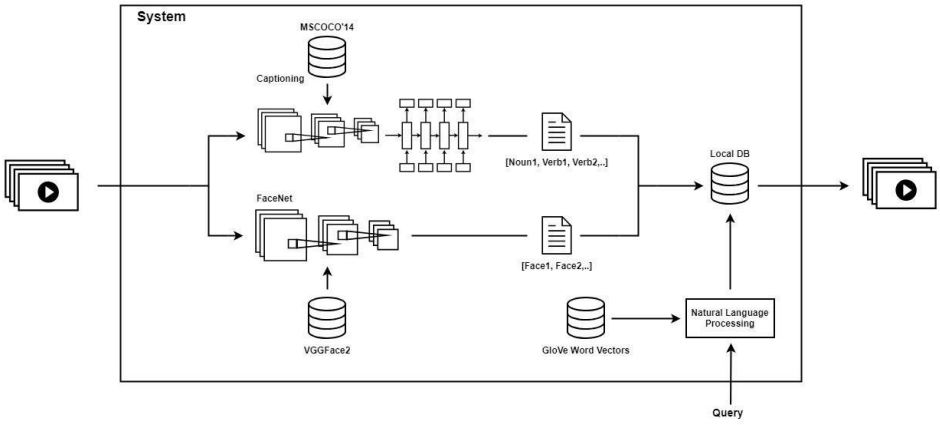

Fig. 1. An overview of the semantic video retrieval system

Fig. 2. Sample video frame. Person recognition output: *'Donald Trump'* and *'Emmanuel Macron'*. Caption Generator output: *'Two men standing next to each other'*.

images were separately collected and not from the set of videos for fair evaluation. Frames extracted from the videos were labeled with occurrences of individuals as well as the action within the frame.We first separately present the results of person recognition and caption generation followed by the combined results of the two modules. As evaluation metrics, we employ the standard precision and recall (and F-measure). Precision refers to the proportion of false positives while recall measures the proportion of true positives. Precision and recall are generally combined into a single measure, the F-measure, which represents the harmonic mean of the two values.

Table I summarizes the precision, recall and F-measure values for person recognition, separately for each of the individuals considered in our study where it can be seen that promising recognition results have been reported. An analysis of the recognition errors revealed that in many cases, the false negatives (missed individuals) were due to faces which were occluded. Likewise, the captioning results on the 50 test videos are summarized in Table II where a high F-measure value of 0.96 is reported.

In addition to the individual results of person recognition and caption generation, we also report the over all retrieval results which include the combination of the outputs of the two modules. Examples of typical queries presented to the system include:*'Vladimir Putin giving an interview'*, *'Donald Trump and Emmanuel Macron sitting together'*, and *'Mark Zuckerberg giving an interview'* etc. For comparison purposes, we reproduce the results of individual modules along with the over all system results in Figure 4. It can be observed that from the view point of natural language based retrieval, the system reports an over all F-measure of 0.85. To provide an insight into the retrieval results, a sample query and the retrieved results are illustrated in Figure 5 where it can be seen that the system correctly fetches the videos relevant to the provided query.

V. Conclusion

This paper presented the findings of a study towards the development of a semantic video retrieval system. More specifically, we focused on the retrieval of videos involving key persons in certain scenarios of interest. We exploited the recent developments in deep learning for object recognition and caption generation to identify persons and the corresponding actions. Identification of individuals was carried out by tuning the FaceNet to our set of images while captions were generated from video frames using the combination of convolutional and recurrent neural networks. Videos were indexed by storing the information on individuals and the respective actions. During the retrieval phase, natural language queries were presented to the system. The concept of word embeddings was employed to find similar words to those in the provided query and the videos containing the queried individuals appearing in a specific scenario were retrieved.

As mentioned earlier, the current findings are the result of a pilot study which is intended to be extended to a comprehensive semantic retrieval system. In our further study, we plan to incorporate objects in addition to persons. Likewise, the textual information appearing in videos can also serve

Fig. 3. Key individuals considered in our study

TABLE I
PERSON RECOGNITION (FACENET) RESULTS ON 50 VIDEOS

Person	Gates	Bajwa	Ghafoor	Trump	Musk	Khan	Macron	Mark	Sundar	Putin
Precision	0.95	0.77	0.87	0.69	0.89	0.83	0.93	0.92	0.97	0.88
Recall	0.90	0.95	0.98	0.97	0.98	0.87	0.93	1	0.97	0.95
F-measure	0.92	0.85	0.92	0.81	0.93	0.85	0.93	0.96	0.97	0.91

TABLE II
RESULTS OF CAPTION GENERATION

Precision	Recall	F Measure
0.93	1	0.96

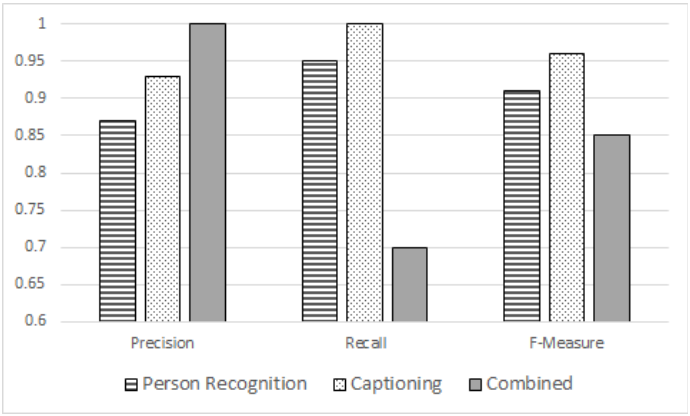

Fig. 4. System performance: FaceNet (Person Recognition), Caption Generation and Over all Retrieval

as a useful index for retrieval purposes. Another interesting modality would be to extract the rich audio information and exploit it to spot keywords or generate textual summarization of the spoken content. In addition to retrieval, the system can be adapted to work on live video streams for applications like generation of alerts (breaking News for instance).

ACKNOWLEDGMENTS

This study is supported by IGNITE, National Technology Fund, Pakistan under grant number ICTRDF/TR&D/2014/35.

REFERENCES

[1] Rachid Benmokhtar and Benoit Huet. An ontology-based evidential framework for video indexing using high-level multimodal fusion. *Multimedia Tools and Applications*, 73(2):663–689, 2014.

[2] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[3] Andreas Eitel, Jost Tobias Springenberg, Luciano Spinello, Martin Riedmiller, and Wolfram Burgard. Multimodal deep learning for robust rgb-d object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 681–687. IEEE, 2015.

[4] Myron Flickner, Harpreet Sawhney, Wayne Niblack, Jonathan Ashley, Qian Huang, Byron Dom, Monika Gorkani, Jim Hafner, Denis Lee, Dragutin Petkovic, et al. The qbic system. *IEEE computer*, 28(9):23–32, 1995.

[5] Lianli Gao, Zhao Guo, Hanwang Zhang, Xing Xu, and Heng Tao Shen. Video captioning with attention-based lstm and semantic consistency. *IEEE Transactions on Multimedia*, 19(9):2045–2055, 2017.

[6] Zhao Guo, Lianli Gao, Jingkuan Song, Xing Xu, Jie Shao, and Heng Tao Shen. Attention-based lstm with semantic consistency for videos captioning. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 357–361. ACM, 2016.

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[8] Saira Jabeen, Gulraiz Khan, Humza Naveed, Zeeshan Khan, and Usman Ghani Khan. Video retrieval system using parallel multi-class recurrent neural network based on video description. In *2018 14th International Conference on Emerging Technologies (ICET)*, pages 1–6. IEEE, 2018.

[9] Ming Liang and Xiaolin Hu. Recurrent convolutional neural network for object recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3367–3375, 2015.

[10] Venice Erin Liong, Jiwen Lu, Yap-Peng Tan, and Jie Zhou. Deep video hashing. *IEEE Transactions on Multimedia*, 19(6):1209–1219, 2017.

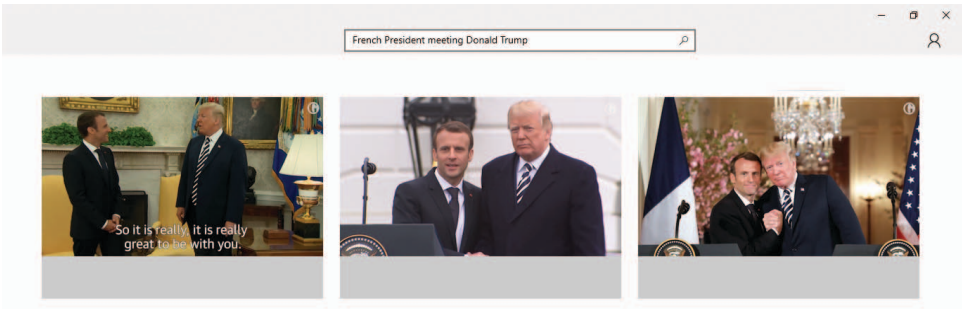

Fig. 5. Sample retrieval results for the query: *'French President meeting Donald Trump'*

[11] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[12] Markus Mühling, Manja Meister, Nikolaus Korfhage, Jörg Wehling, Angelika Hörth, Ralph Ewerth, and Bernd Freisleben. Content-based video retrieval in historical collections of the german broadcasting archive. In *International Conference on Theory and Practice of Digital Libraries*, pages 67–78. Springer, 2016.

[13] Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Naokazu Yokoya. Video summarization using deep semantic features. In *Asian Conference on Computer Vision*, pages 361–377. Springer, 2016.

[14] Alex Pentland, Rosalind W Picard, and Stan Sclaroff. Photobook: Content-based manipulation of image databases. *International journal of computer vision*, 18(3):233–254, 1996.

[15] Haifeng Qi, Jing Li, Qiang Wu, Wenbo Wan, and Jiande Sun. A 3d-cnn based video hashing method. In *Tenth International Conference on Digital Image Processing (ICDIP 2018)*, volume 10806, page 1080644. International Society for Optics and Photonics, 2018.

[16] Yogesh Singh Rawat, Aayush Rana, Praveen Tirupattur, and Mubarak Shah. Action and object detection for trecvid. 2018.

[17] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[18] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[19] Luca Rossetto, Ivan Giangreco, Heiko Schuldt, Stéphane Dupont, Omar Seddati, Metin Sezgin, and Yusuf Sahillioğlu. Imotiona content-based video retrieval engine. In *International Conference on Multimedia Modeling*, pages 255–260. Springer, 2015.

[20] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[21] John R Smith and Shih-Fu Chang. Visualseek: a fully automated content-based image query system. In *ACM multimedia*, volume 96, pages 87–98. Citeseer, 1996.

[22] Elena Stringa, Paul Meylemans, and João GM Gonçalves. Image retrieval by example: Techniques and demonstrations. In *23rd ESARDA Symposium on Safeguards and Nuclear Material Management*, volume 2, 2001.

[23] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[24] Maria Tzelepi and Anastasios Tefas. Deep convolutional learning for content based image retrieval. *Neurocomputing*, 275:2467–2478, 2018.

[25] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*, 2014.

[26] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):652–663, 2017.

[27] Ji Wan, Dayong Wang, Steven Chu Hong Hoi, Pengcheng Wu, Jianke Zhu, Yongdong Zhang, and Jintao Li. Deep learning for content-based image retrieval: A comprehensive study. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 157–166. ACM, 2014.

[28] Wei Wang, Xiaoyan Yang, Beng Chin Ooi, Dongxiang Zhang, and Yueting Zhuang. Effective deep learning-based multi-modal retrieval. *The VLDB JournalThe International Journal on Very Large Data Bases*, 25(1):79–101, 2016.

[29] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.