Instituto Tecnologico y de Estudios Superiores de Monterrey

Monterrey Campus

School of Engineering and Sciences

Self-supervised Learning of Human Action Representation Using Attention Mechanisms for Video Retrieval

A thesis presented by

Jesús Andrés Portillo Quintero

Submitted to the
School of Engineering and Sciences
in partial fulfillment of the requirements for the degree of

Master of Science

in

Computer Science

Monterrey, Nuevo León, June, 2020

Instituto Tecnologico y de Estudios Superiores de Monterrey

Campus Monterrey

The committee members, hereby, certify that have read the thesis presented by Jesús Andrés Portillo Quintero and that it is fully adequate in scope and quality as a partial requirement for the degree of Master of Science in Computer Sciences.

Hugo Terashima Marín
Instituto Tecnológico y de Estudios Superiores de Monterrey
Principal Advisor

To be defined
First Committee Member's institution
Committee Member

To be defined
Second Committee Member's institution
Committee Member

Rubén Morales Menéndez
Associate Dean of Graduate Studies
School of Engineering and Sciences

Monterrey, Nuevo León, June, 2020

Declaration of Authorship

I, Jesús Andrés Portillo Quintero, declare that this thesis titled, Self-supervised Learning of Human Action Representation Using Attention Mechanisms for Video Retrieval and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Jesús Andrés Portillo Quintero
Monterrey, Nuevo León, June, 2020

©2020 by Jesús Andrés Portillo Quintero
All Rights Reserved

Dedication

To my grandmother.

Acknowledgements

I would like to express my gratitude towards my colleagues and teachers, whom without their assistance this work would be impossible. I also would like to thank Tecnológico de Monterrey and CONACyT for the invitation to this program and financial support.

Self-supervised Learning of Human Action Representation Using Attention Mechanisms for Video Retrieval

by

Jesús Andrés Portillo Quintero

Abstract

Information Retrieval systems aided by advancements in Computer Vision have made Text to Video Retrieval systems an active research area. In order to retrieve video from a text query it is needed to transform both modalities into a common representation. This representation can be learned using a Deep Neural Network (DNN) framework, in which the repository of videos and the text query can be processed in order to obtain the most relevant features of them. This work proposes two complementary approaches for text-video representation learning. First, implement an attention-based encoder for either text or video, as the attention mechanism has proven successful for other DNN models. Second, this work seeks to train the DNN in a self-supervised manner as this method has been proven to find rich and sparse representations. This document represents a thesis project for the degree of Master in Computer Science from Instituto Tecnológico y de Estudios Superiores de Monterrey. This is still a work in development, a phase of exploration and replication of similar works has been achieved and the author of this work is preparing the framework in order to prove the first of the above approaches.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

We live in the Era of Information, thousands of gigabytes of are generated each day, but if it can't be organized and accessed there is little benefit. Information Retrieval systems seek to increase the usability of available information from a repository for the benefit of the user, web browsers such as Bing and Google are prominent examples. Nonetheless, it has been stated that the development of this frameworks is far from over [1], especially the IR systems that handle multimedia files, for instance sound, images and video. Multimedia Information Retrieval is the field concerned with the development of tools for search, rank and retrieval of relevant multimedia files given a query. There are several applications for this kind of systems such as music identification, searching medical imaging files, artwork retrieval, text to video retrieval, among others [37].

Text to video retrieval has become an active research field due to its relationships with human activity recognition [29, 35]. It relies on the translation of video containing human actions and semantics of a text query into a common representation, a set of shared discriminative features, that can be compared for similarity measurement [12]. With the advent of Deep Learning several methods for human action representation have been developed for both images and video [7, 22]. Recent representation learning methods have relied on Attention Mechanisms (AM) under Self-supervised (SS) training regimes [9, 39], however little research has actively tried to make a retrieval system using these two paradigms.

Attention mechanisms proliferated from the Natural Language Processing (NLP) field. One of its major breakthroughs has been the development of the Transformer architecture [41], this model is based on a novel self-attention mechanism that learns what parts of an input sequence to use depending on context, giving it a sense of focus. Variants of this seminal work have resulted into transformer-based models trained on a large text corpus in a self-supervised manner [11, 27].

Self supervision refers to a training regime in which a Neural Network model is trained to perform a proxy task which relies on a transformation of the input data to avoid the need of labels in order to enrich the internal representation of the model, thus enhancing its performance [10].

Considering these two recent advancements, we propose to create a text to video retrieval architecture that uses both paradigms. Firstly, the different implementations of retrieval architectures will be explored to identify the segments where an AM can be implemented and test whether said architecture can be used for a pretext task. Then an attention-based model will be

implemented into the selected existing network and be assessed for improvement. Afterwards, the same model is planned to be subject to a self-supervised training regime.

1.1 Problem Statement

Retrieval systems for video have become an active field of research given the amount of videos that are uploaded to the internet every day [29]. Video retrieval consists of the search of video files on a database that are most similar to a query which can either be image, text or other media. Commonly, text to video retrieval relies on metadata cues which denote the content of the media, but it has been demonstrated that this hints are not enough for finding relevant matches to the queries in the plethora of information available [35], especially when there is a presence of dense semantics in the media such as different object instances, quick movements, geographical information and human actions.

For videos that explicitly include human figures or content, a clear understanding of human action is needed. Human action understanding is the body of knowledge that studies the information of what we call actions into its subcomponents such as movement and intention [5]. In the field of computer vision, action recognition is concerned with defining the mechanisms to extract the most relevant features of the media, images or video, and aggregate them into a predefined format. This format is called a representation, a common topic in the Machine Learning field as it relies on the definition of characteristics or features that best describe an object, problem or dataset. Video representations can be harnessed to perform several tasks such as captioning, object tracking, similarity comparison, retrieval, among others which have real life applications for instance intelligent surveillance, autonomous driving, social media filters, person identification, among others. [15].

Representation learning is a research field on itself as it is a fundamental topic in machine learning research [4]. It is concerned with transformation that a data source (i.e. media) is subject to in order to find the most discriminative cues or features. Representations used in retrieval tasks follow the form of dual encoding networks [12, 24]. Where a visual encoder model transforms the video data and a textual encoder transforms the text data into a common representation, hence resulting into a shared text-video representation. There are multiple pathways to formulate a representations, but there is a common consensus in that a good representation in the context of action recognition must be easy to compute, provide description for a sufficiently large class of actions, reflect the similarity between two like actions, and be robust to various variations [17].

Originally representations were hand made through feature engineering where emergent patterns of data were analyzed and processed as features. For visual representations Color Histograms, Scale Invariant Transforms, among other statistical methods were used [35]. Recently, with the advent of Deep Learning (DL) methods the representation of visual cues has been learned rather than engineered. Since the debut of the Neural Network (NN) classifier Alexnet in 2012 in the Imagenet competence, several Deep Learning models have been developed for image related tasks such as classification, text generation, segmentation, object bound boxing, among others [45]. In the case of DL models for video, they have been mostly extensions of said methods that comply with the extra temporal dimension, such as 3D convolutional networks [7], visual Long-Short Term Memory (LSTM) [44] and more recently

Attention-based models [39].

Attention is a novel Deep Learning mechanism, originally proposed for text translation tasks, it provides a method for bringing focus and context into a model [3]. This component has seen great success in applications of sequence representation learning in the field of Natural Language Processing. Transformers, an architecture that implements self-attention, an adaptation of the original attention mechanism, has taken NLP models to unparalleled performance, mostly due to the success of its variants (BERT [11], RoBERTa [27]) and its capabilities of learning in a self-supervised way.

There has been some implementations on image related tasks such as UNITER [9] and on videos with VideoBERT [39], with both using an implementation of a transformer and its training has used some form of pretext task. A pretext or proxy task is a form of self-supervised learning in which a model uses a transformation on the very main dataset as a label. Both models have shown good results based on the quality of its representation, but they have fallen short on the applications of video retrieval because of the computational expensiveness of similarity search on a database with their architecture [36].

A retrieval system that uses the advantages given by attention mechanisms should be able to generate the representations of video and text needed for efficient similarity measurement. Also the retrieval system should use a form of self-supervision for training the representation as the annotated datasets available for video are not as plenty as ones for images [28].

1.2 Hypothesis

We argue that implementing Attention-based models into a dual encoder architecture trained under a Self-supervised regime can translate into learning a rich common text-video representation that can be used for Video Retrieval and present a prominent performance enhancement with respect to other similar approaches in the literature.

- What are the most appropriate dimensions and characteristics of a dual encoder Neural Network to learn a common text-video representation?
- What is the most appropriate dataset for network training? How will its performance be validated?
- How the implementation of an attention-based model in either text encoder or video encoder will impact retrieval performance?
- What kind of pretext tasks can be performed with the resulting architecture that enable a form of self-supervised learning?

1.3 Objectives

The main objective of this research work is to develop a Video Retrieval framework that uses an attention-based model to encode either visual or textual information to learn a common text-video representation under a self-supervised training regime. In order to verify whether

the hypothesis is correct or requires a restatement several objectives have been laid out that are in line with the aforementioned research questions.

- Explore and replicate dual encoder implementations by other authors in order to find possible modifications.
- Implement an attention-based model into an existing architecture and measure its performance to test whether its success in NLP can be replicated in Video Retrieval.
- Define a pretext task and train the model accordingly, to test whether the Self-supervised learning enhanced the representation performance.

1.4 Solution Overview

This document is concerned with the development of a model for similarity comparison between video and text using a dual encoder Neural Network which can serve for retrieving videos from a database.

Overall, this work intends to project text and video into a common space using both Natural Language Processing models and Convolutional Neural Networks for video. This with the finality of comparing several videos to a query and retrieve the nearest semantically.

Other Video Retrieval architectures are in use nowadays, but require of metadata for explaining its content. On the contrary, the model proposed in this work attempts to encode videos into a representation that can be directly compared with other text-video encodings, thus eliminating the need for metadata or any kind of hint to relate a given query to a specific video.

The model is trained with a set of video and caption pairs whose representations are similar using a distance metric. A fundamental property of the learned representation is that it must encode semantically similar videos or captions near, while distancing unrelated pairs. The distancing learning can be enhanced by Pair Mining, which consists on the selection per each item on the dataset the most contrary data points and make the Neural Network locate the two pairs separately on their respective representations.

1.5 Contributions

This dissertation takes intends to join several novel archetypes that are present on tangential fields such as NLP and Video Processing. Nonetheless, to the best of our knowledge, the approach presented in this document has not been implemented. The most important contribution, at this research phase, is the utilization of text similarity as proxy for Negative Pair Mining. On more detail, the contributions that are present (or will be present) in this document are:

Analysis of sentence embeddings for Representation Learning

The text encoder on other works is comprised by an aggregation of word-level representation. On the other hand, this work analyses the utility of sentence-level representation.

Use of text similarity as Negative Pairing proxy

Representations for similarity comparison are greatly benefited from Negative Pair Mining, a procedure on which a data point is contrasted with its most contrary, i.e. most distant, on a dataset. This procedure has to be repeated each time the representation is updated which increases training time. A possible way to avoid consecutive mining is to use the text representation as an absolute metric. Negative Pairs would be mined only once and the visual encoder will be the one that has to adapt to the text representation space.

Training on a Self-supervised regime

Models trained on an unsupervised manner tend to perform better than their fully supervised counterpart. It is expected of this work to prove the feasibility of this training regime on the Video Retrieval task.

1.6 Thesis Organization

The ensuing chapters of this dissertation detail technical information relevant to this investigation. Chapter 2 presents the background and state of the art of the topics related. An overview of exploration and proven models are shown in Chapter 3. In Chapter 4, the methodology used during this research is described. Chapter 5 presents the basic model used and its technical and implementation details. This chapter serves as a preamble of the two subsequent chapters. Chapter 6 will present a more advanced model based on the learnings from the previous chapter. Results will be compared in Chapter 7. Finally, Chapter 8 will provide conclusions derived from the investigation.

Chapter 2

Background and Related Work

The theoretical foundations of this project proposal begin with the description of basic Computer Vision concepts. Next, Machine Learning methods are introduced and its influence on novel architectures for feature extraction and Human Action Recognition. Then moves on to explain the field of content based retrieval of media. To finalize with current approaches for video retrieval.

2.1 Computer Vision

Computer vision is the ability of a computer to recognize and interpret the content of an image or video. Is a process through which visual sensation is transformed into visual perception. During this process, a computer receives visual data, analyzes it, and makes some decision about that data [23]. This field relies on a plethora of mechanisms that allow for the interpretation of images. For the purpose of this document we will cover selected aspects of it for further reference.

2.1.1 Convolutions

Convolution is a fundamental operation in image processing. Is the product of each pixel value and change it in some way. To apply this mathematical operation, we use another matrix called a kernel. For example in Figure 2.1 the input image is denoted by the white square, while the kernel is gray square. For each pixel in the image, we take the kernel and place it on top such that the center of the kernel coincides with the pixel under consideration. We then multiply each value in the kernel matrix with the corresponding values in the image, and then sum it up. This is the new value that will be substituted in this position in the output image [19].

For example, edge detection can be done by using a kernel, also known as filter, to identify horizontal and vertical changes of color (Figure 2.2).

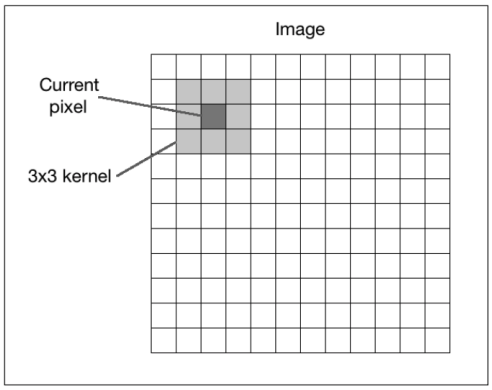

Figure 2.1: A convolution procedure.

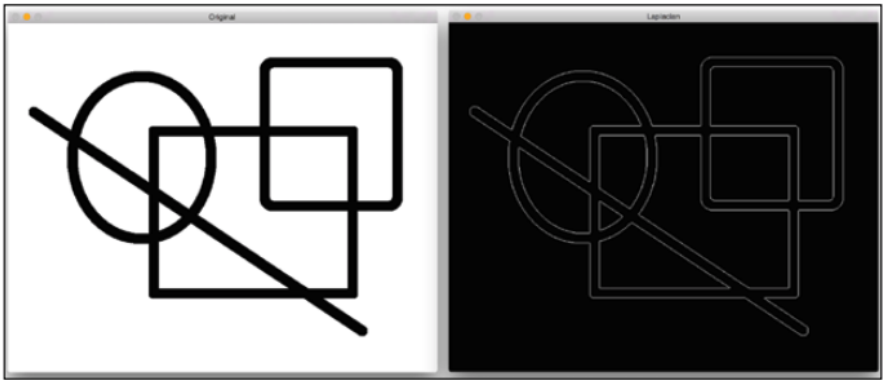

Figure 2.2: Edge detection filter.

2.1.2 Object Tracking

Object tracking refers to the problem of using sensor measurements to determine the location, path and characteristics of objects of interest [8]. The typical objectives of object tracking are the determination of the number of objects, their identities and their states, such as positions, velocities, among others. An example is observing the trajectory of a subject walking, see Figure 2.3. According to Murugavel it includes [32]:

- Motion detection: Has as only purpose detect if there is in fact a motion event on a specific time.
- Object localization: Focuses attention to a region of interest in the image.
- Motion segmentation: Images are segmented into regions corresponding to different moving objects.
- Three-dimensional shape from motion: Also called structure from motion, intends to supply the notion of depth.

Figure 2.3: Subject tracking.

2.1.3 Classification

Classification is related to the methodological organization of image or video data into a group of related media. Usually the comparison is made on a numerical basis, features from a file are converted into a magnitude which can then be compared against other feature mappings using a comparison distance [34]. For example, assuming images of color blue have an associated value of 2, two images that show a blue sky should be close to that rank (see Figure 2.4). Of course there are several more features an image could have, the same procedure can be run for a set of features in the form of a matrix or vector. In order to group several distance measures, a Nearest Neighbor clustering approach is a common solution.

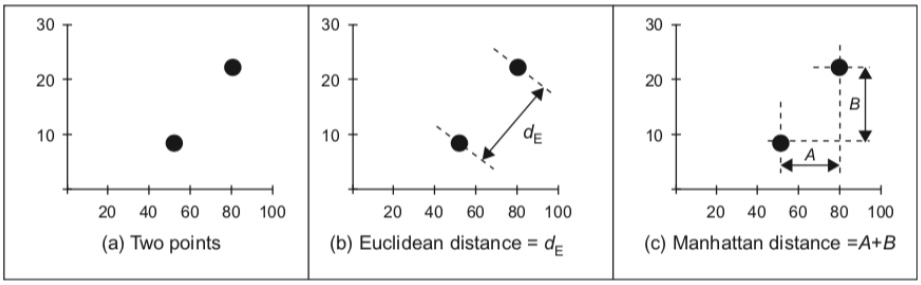

Figure 2.4: Euclidean and Manhattan distance.

Nixon et al. proposes the following classification measures [34]:

- Euclidean distance.
- Manhattan distance.
- Mahalonobis distance.
- Bhattacharyya distance, or cosine distance.

2.2 Machine Learning

Machine Learning is a scientific discipline focused on the development of algorithms that allow computer systems to acquire knowledge about the world from empirical data and generate behaviours useful for mankind [33]. The advantage of this kind of algorithms is that the inner workings are not directly programmed, but learned from supplied information. This capability can be harnessed in order to develop this research project. There are several Machine Algorithms, some will be explained in more detail in further sections.

2.2.1 Nearest Neighbor Instance Learning

An instance is a subset of a dataset, variations of a population. Nearest Neighbor is an algorithm that uses using a type of distance measurement between its dimensions [14]. The goal is to obtain a set of regions inside a plane that can classify new data as a certain instance.

2.2.2 K-means Clustering

Like Nearest Neighbor, the K-means algorithm has as an objective to define the instances of a given unlabeled dataset [14]. This means that classification instances are not given, but have to be learned from the data features.

2.2.3 Support Vector Machine

A Support Vector Machine is a classification algorithm in which the objective is to define a hyperplane that can divide linearly classes on a given dataset [14]. It strives to maximize the margin between a class subset and the separator. When data is not linearly separable, a kernel transformation is used in order to change the relationship between dimensions of data.

2.2.4 Artificial Neural Networks

Artificial neural networks (ANNs) is a computational graph that is inspired from a simplified understanding of the human brain. They are not directly programmed, but they learn from input data. ANNs are composed of interconnected nodes (see Figure 2.5) that contain a linear combination of weights, inputs from other nodes and biases (a scalar term). This components change in the training process and serve as arguments for an activation function, its value serves as output of a node. They can be trained in a supervised or unsupervised manner. In a

supervised ANN, the network is trained by providing matched input and output data samples, with the intention of getting the ANN to provide a desired output for a given input. Layers are the differentiating unit in ANNs, its outputs are the inputs of other layers. They can be described in three basic categories: input layer, hidden layers and output layer. The first is the nexus between training data and the computational graph. The second refers to the layers that are not reachable from outside, hence hidden, in which most of the computation is done. The third is the output layer, which describes the result of the calculation given by the network. In Figure 2.5 a two-layer ANN is shown, notice it has four layers, but the input and output layers are not considered.

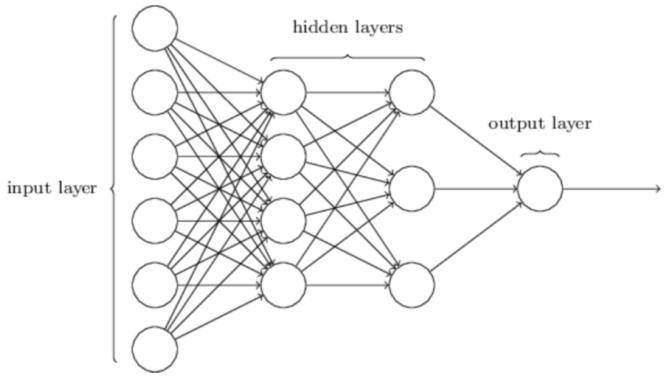

Figure 2.5: A simple two-layer ANN.

2.2.5 Deep Learning

In recent years the term Deep Learning was coined as a special case of an ANN. A Deep Neural Net (DNN) follows the same graph design but has several layers that increase its capabilities. This advantage comes at a two-fold cost: computational complexity and data requirements [13].

2.2.6 Common Artificial Neural Networks Architectures

Several architectures have been made since the inception of ANN. This as a result of years of research, that have tested several patterns in the construction of graphs that are particularly proficient at specific tasks :

- Residual Neural Network
- Recurrent Neural Network
- Long-Short Term Memory
- Recursive Neural Network
- Transformer Network
- Generative Adversarial Network

2.2.7 Feature Embedding

ANN allow for unsupervised learning of features, given there is enough information (as seen in Figure 2.6). Embeddings are low-dimensional representations. This allows for the automatic selection of features, visualization and semantic hashing [6].

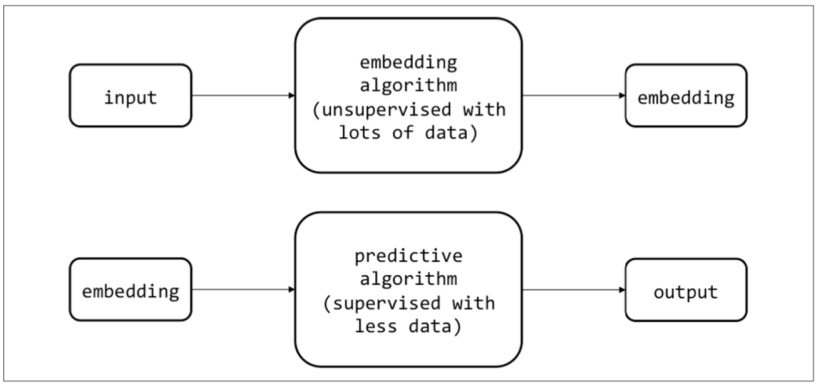

Figure 2.6: Comparison of feature learning approaches [6].

2.2.8 Representation Learning

Representation Learning at its core is the research of methods in which raw data can be transformed to work with a computational model. Bengio et al. state that representations are vector or tensors that store the features of an entity such as image, audio, word or sentence [4].

2.2.9 Attention Mechanism

An Attention Mechanism is a Neural Network component that has been used recently for sequence representation learning. It was first used in translation tasks, but rapidly spread to other research areas [3]. This mechanisms are designed to be able to choose which inputs to use, either part or present, in a sequence. A variant called self-attention was introduced within the Transformer model [41]. Is a mechanism relating different positions of a single sequence in order to compute a representation of the same sequence. The power of this model was its parallelization capabilities, as attention modules can be calculated regardless of the sequence order, which in the case of other recurrent models this is impossible.

A recently studied problem is how to create a joint representation of different data sources, or modalities [15]. A known method for addressing this problem is projecting the heterogeneous features of the data sources into a common subspace, where the multimodal data with similar semantics will be represented by similar vectors [35].

A multimodal representation can be useful for some applications regarding valuable translation of semantics for instance describing images or videos, object classification, image or video retrieval, among others. A good representation must be easy to compute, provide description for a sufficiently large class of instances, reflect the similarity between two like classes, and be robust to various variations [17].

In the past decade there has been a shift from hand-designed representations for specific applications to data-driven based on Deep Neural Networks. Due to the multilayer nature of Neural Networks each successive layer is hypothesized to represent the data in a more abstract way, hence it is common to use the final or penultimate neural layers as a form of data representation [4].

2.3 Action recognition

Action recognition in computer vision is the research area concerned with the interpretation of movements, pose and spatial data that a humanoid figure can present. Human action recognition has a wide range of applications, such as intelligent video surveillance and environmental home monitoring, video storage and retrieval, intelligent human–machine interfaces, and identity recognition [45]. It is considered a challenging problem as the efficacy of a system of this sorts relies on how robust human action modeling and feature representation methods are designed. Is especially hard on video since the problem of feature representation is extended from two-dimensional space to three-dimensional space-time.

2.4 Self-supervised learning

Self-supervised learning technique where the training data is autonomously labelled by an algorithm that exploits characteristics of a dataset. A self-supervised task, also known as pretext task or proxy task, is performed with the intent of learning a representation state that can carry good semantic or structural meanings and can be beneficial to downstream tasks. For instance, grayscale images, image inpainting, image jigsaw puzzle, etc. [20] are problems where an algorithm can easily obtain a pseudolabel (the original state of the image) and apply a transformation for the network to solve it.

2.5 Multimedia Information Retrieval Systems

Multimedia Information Retrieval Systems (MIRS) are tools used for information accessibility. Are algorithms in charge of extracting useful files based on a request. Kambau and Hasibuan [21] propose that images, video and audio are the three main types of media that can be served by these systems. Retrieval of each of them has its difficulties, the authors also propose three main approaches in which the task can be done:

- Content-based: This method uses the innate characteristics of the media to summarize its main features, in order to facilitate the search for other files with similar characteristics. The content-based MIRS for each media type are called:
  - Content-Based Image Retrieval.
  - Content-Based Video Retrieval.
  - Content-Based Audio Retrieval.

The authors argue that this technology was the first being developed for the purpose of information retrieval. They also state it is the less computational intensive.

- Context-based: The request is a combination of search technologies and knowledge about the user being. In other words the algorithm is assisted by multi-modal information input. The MIRS for each media type are called:
  - Context-Based Image Retrieval.
  - Context-Based Video Retrieval.
  - Context-Based Audio Retrieval.

  The authors argue that this technology was the second being developed for the purpose of information retrieval. They also state it is more computational intensive than content-based algorithms.

- Concept-based: This method requires an understanding of the meaning of the query. Concepts have to be represented in some way, the first approaches were with tags and weights. The MIRS for each media type are called:
  - Concept-Based Image Retrieval.
  - Concept-Based Video Retrieval.
  - Concept-Based Audio Retrieval.

  This is the most recent development on MIRS due to the complexity of knowledge representation for the main types of media. Novel architectures such as WordNet and Bag of Words have created a tractable representation of semantics [21]. This project will be developed mainly in the field of Concept-Based Video Retrieval.

2.6 Triplet Loss

Triplet Loss, Max Ranking Loss or Hinge Loss presented by Shroff et al. is a function that forces similar instances of a subject to be close in an embedding space [38], as well as dissimilar or unrelated subjects to be clustered apart. This function has as inputs a triplet of embeddings (a, p, n) as well as a $margin$ hyperparameter:

$$\mathcal{L} = \max(d(a, p) - d(a, n) + margin, 0) \tag{2.1}$$

The goal of this function is to minimize it, this moves $d(a, p)$ close to zero and makes $d(a, n)$ an element that can drive the sum to a negative value, thus effectively driving the loss to zero. Based on this behavior there are three major scenarios for triplets:

- Easy triplet: The loss is zero because $d(a, p) + margin < d(a, n)$.
- Hard triplet: Triplet where the negative is positioned closer to the anchor than the positive, in other words $d(a, p) < d(a, n)$.

- Semi-hard triplet: Triplet where the positive is closer to the anchor than negative, but the difference between them is not greater than the margin, thus the loss is positive: $d(a,p) < d(a,n) < d(a,p) + margin$.

As the previous cases depend entirely on the position of negatives they can also be defined as easy, hard or semi-hard negatives. The kind of negatives used for training have a direct impact on the quality of representation [18], consequentially, a sample strategy is needed to sample the best triplets suited for training.

2.6.1 Negative Mining

Negative Mining is a procedure of actively sampling training data points to be used on a Triplet Loss or similar functions. There are two major approaches that are implemented in literature:

- Offline Mining: Identify the most suitable triplets from all the training set at the beginning of each training procedure. This method produces more valid triplets, but it is computationally expensive.

- Online Mining: During batch processing compare items only on the present batch for triplet sampling. This method is used on most implementations.

2.6.2 Triplet Loss on Video Retrieval

Video Retrieval architectures rely on a modification of the Triplet Loss to accommodate related videos and text close on an embedding space. A generalization of said function takes into account the negative representation of both text and video to be contrasted with a positive data point. Given that the datasets for Video Retrieval contains video and caption pairs, the loss should be calculated twice in order to provide a mean to optimize both video and text representations. This is expressed as a sum of two Equations 2.1 into Equation 2.2 [25, 28].

$$\sum_{i \in \mathcal{B}} \sum_{j \in \mathcal{N}(i)} max(0, margin + d_{i,j} - d_{i,i}) + max(0, margin + d_{j,i} - d_{i,i}) \tag{2.2}$$

On Equation 2.2 positive pair i from video-caption set $\mathcal{B}$ are contrasted with their respective negatives j from $\mathcal{N}(i)$, where $\mathcal{N}(i)$ is the negative pair sampler for data point i.

2.7 Related Work

This section contains a brief description of articles that are considered relevant for the research problem, as they are considered novel on their methodology, approach, or impact.

2.7.1 Use What You Have: Video Retrieval Using Representation From Collaborative Experts

In this article Liu et al. propose an architecture based on ”collaborative experts”. This approach uses pre-trained semantic embeddings (”experts”) such as Automatic Speech Recognition (ASR) and Optical Character Recognition (OCR) in order to augment the amount of information a video may provide. In other words uses audio, text shown in the video, scene interpretation, among others to increase the descriptive power of the media, which is then converted into a common feature embedding using a Transformer Network [26]. The main contribution of this work is that reduces the annotation scarcity, is an ensemble of classificators. As can be seen in Figure 2.7, the retrieval task is done by encoding a free form query into a vector of the same dimensions as the multi-modal embedding. This is the first candidate for result replication.

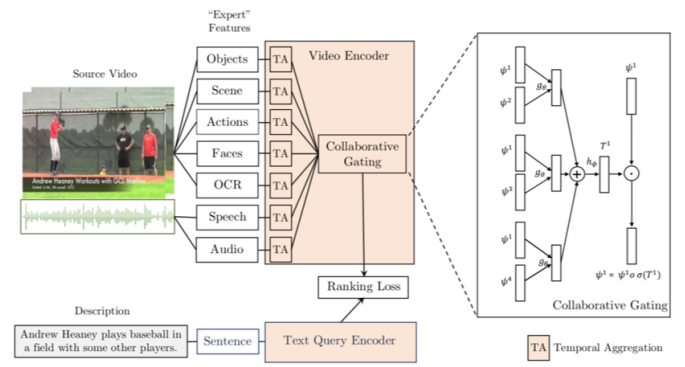

Figure 2.7: Architecture proposed by Liu et al. [26].

2.7.2 HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips

This article is mostly concerned on the process to build a dataset that is completely sentence annotated. This means the dataset is useful for training dense-captioning videos and, most important, text based video retrieval. In the first part Miech et al. propose a method in which it is possible to obtain reasonably good annotations from tutorial videos. Given that most tutorial videos follow a step-by-step scheme and are narrated in the process. This fact, along Automatic Speech Recognition system, allow for a more rich and automated annotation [29]. This methodology may serve in future for creating a dataset of suspicious or crime activities, one idea is to use news video clips in order to obtain the crime section and its narration by the news anchor. This to generate a video-text pair for training. Another contribution made by Miech et al. is that the embedding can be generalized to other contexts. The embedding they trained achieved state-of-the art results on instructional video datasets, but they claim the model also works on other video description databases well.

2.7.3 W2VV++: Fully deep learning for ad-hoc video search

Li et al. propose the W2VV++ model, a dual encoder Neural Network, for Ad-hoc video search. The architecture of the model is based on a Word2vec pooled with a recurrent Neural Network for text encoding and a pretrained Convolutional Neural Network for visual feature extraction [24]. Triplet loss is used as loss function, this formula has proven to be useful for learning sparse representation on which similar concepts are closer. This model can be considered simple in comparison with the work of Liu et al. [26]. The relative simplicity of this architecture makes it prone to modifications as there are few interconnected components that may hinder the implementation of new ones.

Tan et al. propose a method for not only retrieving a video, but retrieving the moment in which the sentence applies [40]. They propose a weakly supervised model, a type of learning in which the annotations are not completely accurate, called wMAN. The model is comprised by a Frame-by-Word module and a World-Conditioned Visual Graph. The first exploits the similarity of scores on the frames of a moment. The second constructs a graph that updates its nodes with the semantincs of each passing frame, until a phrase is constructed.

2.8 Summary

In this chapter the theoretical foundations of this research work were laid. The fields of Machine Learning provide with the algorithms necessary for Representation Learning, as well as the basic components used in a Neural Network. Computer Vision is another knowledge area on which this work relies heavily, it is concerned with the analysis of imagery for extraction of relevant objects, colors or movements. Action Recognition is the field for describing and identifying the movements and gestures from a humanoid shape, Computer Vision methods and a good representation are fundamental for its performance. Self-supervised Learning is a regime in which a model is subject to other training set in order to robust the descriptive capabilities of its architecture. The Related Works are articles in which the retrieval of videos is achieved by precomputing a representation of each video file and comparing it against the representation of the input query.

Chapter 3

Solution Model

In order to develop a model that can fulfill the stated objectives we propose to start by exploring and replicating the works from Li et al. [24] and Liu et al. [26]. The W2VV++ made by Li et al. is a lightweight model that can be useful for a proof of concept to familiarize the author with the Video Retrieval task. The Liu et al. model contains more pretrained classificators that aid for the representation solution, this makes it worthy of replication for understanding.

As the first objective relates to the feasibility of using a Attention-based sentence embedder the model and video features provided by Li et al. will be used. As this work, on which we shall start experimentation, has a dual encoder architecture it makes it prone to simple modifications such as implementing a different model for text encoding.

The model will be trained on the MSR-VTT dataset [43] and evaluated using the Recall at k metric which denotes the quality of retrieved items. Using the findings from the modified Li et al. model it is planned to subdue the model to a training regime for a pretext task. It is still unclear which task can be performed with the architecture at hand, but the most prominent candidates for implementation are Video Captioning and Video Alignment.

Chapter 4

Methodology

In order to reach the objectives of this project the following set of activities needs to be accomplished:

1. **Literature review**

   A deeper introspection into the body of knowledge is needed to fully understand the theoretical foundations of this work. This field is constantly being updated with novel approaches, so being up to date is important.

2. **Personal learning**

   This project requires multidisciplinary competences, several fields of Computer Science are taken into account in the most recent works. To follow along, it will be needed to further increase capabilities in Neural Networks principles, computer vision, linear algebra, among others. This will be done by attending online courses and workshops. More learning resources will be considered in the future.

3. **Experiment replication**

   This step consists on finding documented code resources that solve the video retrieval problem. This is with the purpose of familiarizing the candidate with the existing frameworks, benchmarks and practices in the field. Several conferences and competitions make their code available. Initially the candidate will focus on the replication of results from [29].

4. **Architecture design and analysis**

   Once the standard practices and architectures are familiar, the candidate will proceed towards designing its own architecture for video retrieval in accordance to the research objectives. Also a series of tests have to be made in order to verify its performance and capabilities.

5. **Comparison**

   Once a stable design is made, it will be tested along other models to check how well does it stand against similar approaches and justify its contribution to the field and applicability to the context of security.

6. **Report**

   The writing and defense of the research thesis will take place in this step. Also the publication of a research paper is considered.

Chapter 5

Basic Model

The Basic Model proposed is based on the work of Li et al. [25]. It assumes a set of n video clips and caption pairs $\{(V_i, C_i)\}_{i=1}^{n}$. Each pair has its features extracted by a ResNet and Word2Vec into a vector $\mathbf{v} \in \mathbb{R}^{d_v}$ and $\mathbf{c} \in \mathbb{R}^{d_c}$ respectively. The objective of the model is learn a mapping to a common vector of size d by the functions $f : \mathbb{R}^{d_v} \to \mathbb{R}^d$ and $g : \mathbb{R}^{d_c} \to \mathbb{R}^d$, where d_v is the input size if visual features and d_c is the input size of text features. The cosine similarity among inputs should be low when pairs are alike in content and high on the contrary (Equation 5.1).

$$s(V, C) = \frac{f(\mathbf{v}) \cdot g(\mathbf{c})}{\parallel f(\mathbf{v}) \parallel_2 \parallel g(\mathbf{c}) \parallel_2} \tag{5.1}$$

Functions 5.2 and 5.3 are simple Feed Forward Neural Networks, where $W_1^v \in \mathbb{R}^{d \times d_v}$, $W_1^c \in \mathbb{R}^{d \times d_c}$, b_1^v, $b_1^c \in \mathbb{R}^{d \times d}$ are parameters to learn and σ is the sigmoid activation function. The authors proposed a dimension of $d = 4,096$ for the common representation.

Feature extractors used for video are the 2D CNN Resnet-152 [16] and ResNeXt101 [42] pre-trained on ImageNet and the output of each is concatenated into a vector of size d. Captions are stripped by discarding English stop-words, the remaining words are embedded using the a Word2Vec model pretrained with GoogleNews [31].

$$f(\mathbf{v}) = \sigma(W_1^v \mathbf{v} + b_1^v) \tag{5.2}$$
$$f(\mathbf{c}) = \sigma(W_1^c \mathbf{c} + b_1^c) \tag{5.3}$$

The model is trained using a max margin ranking loss. Using the Equation 5.4 it is computed from a batch $\mathcal{B}$ of video-captions pairs $(V_i, C_i)_{\in \mathcal{B}}$, $\mathcal{N}(i)$ refers to a set of unrelated pair for the i data point, $s_{i,i}$ is the cosine similarity score of video V_i and caption C_i, and δ is a fixed margin parameter. The resulting gradient serves to update the model parameters.

$$\sum_{i \in \mathcal{B}} \sum_{j \in \mathcal{N}(i)} max(0, \delta + s_{i,j} - s_{i,i}) + max(0, \delta + s_{j,i} - s_{i,i}) \tag{5.4}$$

Our proposition is to replace the text encoder by a Transformer-based model that can embed an input sequence into a constant length vector $\mathbf{c}$ (Figure 5.1). These kind of models

rely on the Attention Mechanism discussed before, we argue that using a pre-trained model that implements it can achieve a better text representation. Hence, we will use the Sentence-BERT [36], a BERT based model which pools its output sequence into a vector of length d.

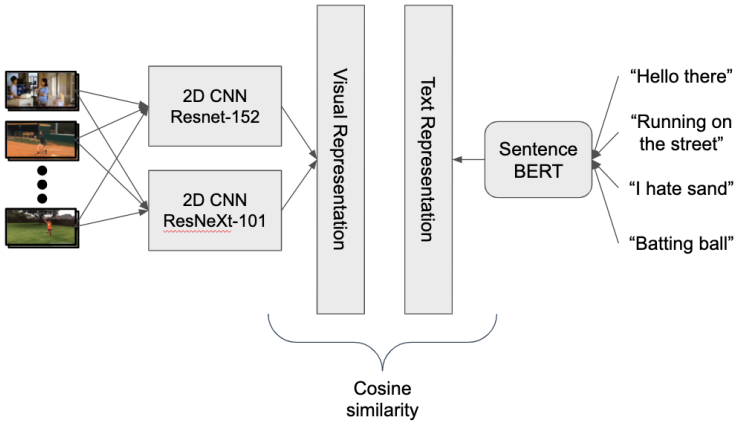

Figure 5.1: Model proposed in this work.

The dataset used for initial training the text-video representation is the TV16-VTT-train dataset [2] which consists on 200 videos with two captions each. This dataset was chosen given the availabilty of precomputed features and ease of experimentation. On a more advanced phase the training set will be the dataset MSR-VTT [43].

The main purpose of this architecture is the retrieval of video files, the standard metric for retrieval is called Recall at k (R@k) which measures the proportion of relevant documents that are in the top k. It is a standard procedure to test the retrieval capabilities of a model at top 1, 5 and 10. In Table 5.1 can be seen the R@k metrics of the work that, to the best of our knowledge, is considered State-of-the-art [30].

Method	Training set	R@1	R@5	R@10
Miech et al. [30]	HowTo100M	6.1	17.3	24.8

Table 5.1: Miech et al. model retrieval results.

5.1 Preliminary Results

As stated on the previous section the dataset of choice was TV16-VTT-train [2]. The basic model was split on a 150/50 proportion for training and testing respectively. Negative pairs were not mined, only assigned to previous data point on the dataset. On Table 5.2 it can be seen that the R@k metrics are close to zero, which is an indicator of bad performance. Given the complexity of implementation of a pair miner, a proxy for negative pair selection was proposed. Instead of comparing each element of the batch against each other o obtain hard data points, the pairs are mined from the comparison of the respective sentences of each data point. In other words, we are using the text embeddings as absolute metrics for negatives

localization, this means that negative mining is not needed at each update of functions $f(\cdot)$ and $g(\cdot)$

Recall	Training Set	R@1	R@5	R@10
Score	TV16-VTT-train	0%	6%	10%

Table 5.2:

As can be seen on Table 5.3 the use of the proxy has increased the performance of the model at most two-fold. This may be and indication that the use of text similarity can be beneficial to also estimate video similarity.

Recall	Training Set	R@1	R@5	R@10
Score	TV16-VTT-train	2%	12%	16%

Table 5.3:

Bibliography

[1] *ICMR '20: Proceedings of the 2020 International Conference on Multimedia Retrieval* (Richland, SC, 2020), International Foundation for Autonomous Agents and Multiagent Systems.

[2] AWAD, G., FISCUS, J., JOY, D., MICHEL, M., SMEATON, A., KRAAIJ, W., ESKEVICH, M., ALY, R., ORDELMAN, R., RITTER, M., ET AL. Trecvid 2016: Evaluating video search, video event detection, localization, and hyperlinking.

[3] BAHDANAU, D., CHO, K. H., AND BENGIO, Y. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (9 2015), International Conference on Learning Representations, ICLR.

[4] BENGIO, Y., COURVILLE, A., AND VINCENT, P. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence 35*, 8 (6 2012), 1798–1828.

[5] BORGES, P. V. K., CONCI, N., AND CAVALLARO, A. Video-based human behavior understanding: A survey. *IEEE Transactions on Circuits and Systems for Video Technology 23*, 11 (2013), 1993–2008.

[6] BUDUMA, N., AND LOCASCIO, N. *Fundamentals of Deep Learning: Designing Next-Generation Machine Intelligence Algorithms*, 1st ed. O'Reilly Media, Inc., 2017.

[7] CARREIRA, J., AND ZISSERMAN, A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 2017-Janua* (5 2017), 4724–4733.

[8] CHALLA, S., MORELANDE, M., MUŠICKI, D., AND EVANS, R. *Fundamentals of Object Tracking*. Cambridge University Press, 2011.

[9] CHEN, Y.-C., LI, L., YU, L., KHOLY, A. E., AHMED, F., GAN, Z., CHENG, Y., AND LIU, J. UNITER: UNiversal Image-TExt Representation Learning.

[10] DAI, A. M., AND LE, Q. V. Semi-supervised Sequence Learning. *Advances in Neural Information Processing Systems 2015-Janua* (11 2015), 3079–3087.

[11] DEVLIN, J., CHANG, M.-W., LEE, K., AND TOUTANOVA, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

[12] DONG, J., LI, X., XU, C., JI, S., HE, Y., YANG, G., AND WANG, X. Dual Encoding for Zero-Example Video Retrieval. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2019-June* (9 2018), 9338–9347.

[13] DUNN, T. Deep learning. *Salem Press Encyclopedia of Science* (2019).

[14] GOLLAPUDI, S., AND LAXMIKANTH, V. *Practical Machine Learning.* Community Experience Distilled. Packt Publishing, 2016.

[15] GUO, W., WANG, J., AND WANG, S. Deep Multimodal Representation Learning: A Survey. *IEEE Access 7* (2019), 63373–63394.

[16] HE, K., ZHANG, X., REN, S., AND SUN, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (12 2016), vol. 2016-Decem, IEEE Computer Society, pp. 770–778.

[17] HERATH, S., HARANDI, M., AND PORIKLI, F. Going deeper into action recognition: A survey. *Image and Vision Computing 60* (4 2017), 4–21.

[18] HERMANS, A., BEYER, L., AND LEIBE, B. In Defense of the Triplet Loss for Person Re-Identification.

[19] HOWSE, J. *OpenCV Computer Vision with Python*. Packt Publishing Ltd, 2013.

[20] JING, L., AND TIAN, Y. Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2 2019), 1–1.

[21] KAMBAU, R. A., AND HASIBUAN, Z. A. Evolution of information retrieval system: Critical review of multimedia information retrieval system based on content, context, and concept. In *2017 11th International Conference on Information Communication Technology and System (ICTS)* (Oct 2017), pp. 91–98.

[22] KRIZHEVSKY, A., SUTSKEVER, I., AND HINTON, G. E. ImageNet classification with deep convolutional neural networks. *Communications of the ACM 60*, 6 (6 2017), 84–90.

[23] LASKY, J. Computer vision. *Salem Press Encyclopedia of Science* (2019).

[24] LI, X. Deep learning for video retrieval by natural language. In *FAT/MM 2019 - Proceedings of the 1st International Workshop on Fairness, Accountability, and Transparency in MultiMedia, co-located with MM 2019* (New York, New York, USA, 10 2019), Association for Computing Machinery, Inc, pp. 2–3.

[25] LI, X., XU, C., YANG, G., CHEN, Z., AND DONG, J. W2Vv++: Fully deep learning for ad-hoc video search. In *MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia* (10 2019), Association for Computing Machinery, Inc, pp. 1786–1794.

[26] LIU, Y., ALBANIE, S., NAGRANI, A., AND ZISSERMAN, A. Use what you have: Video retrieval using representations from collaborative experts, 2019.

[27] LIU, Y., OTT, M., GOYAL, N., DU, J., JOSHI, M., CHEN, D., LEVY, O., LEWIS, M., ZETTLEMOYER, L., AND STOYANOV, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach.

[28] MIECH, A., ALAYRAC, J.-B., SMAIRA, L., LAPTEV, I., SIVIC, J., AND ZISSERMAN, A. End-to-End Learning of Visual Representations from Uncurated Instructional Videos.

[29] MIECH, A., LAPTEV, I., AND SIVIC, J. Learning a text-video embedding from incomplete and heterogeneous data. *CoRR abs/1804.02516* (2018).

[30] MIECH, A., ZHUKOV, D., ALAYRAC, J.-B., TAPASWI, M., LAPTEV, I., AND SIVIC, J. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. *Proceedings of the IEEE International Conference on Computer Vision 2019-Octob* (6 2019), 2630–2640.

[31] MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G., AND DEAN, J. Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems* (10 2013).

[32] MURUGAVEL, M. Multiple object tracking algorithms, December 2019.

[33] NILSSON, N. J. *Introduction to machine learning. An early draft of a proposed textbook*. Stanford, 1996.

[34] NIXON, M., AND AGUADO, A. *Feature extraction & image processing for computer vision*, 3 ed. Academic Press, 2012.

[35] RASIWASIA, N., COSTA PEREIRA, J., COVIELLO, E., DOYLE, G., LANCKRIET, G. R., LEVY, R., AND VASCONCELOS, N. A new approach to cross-modal multimedia retrieval. In *MM'10 - Proceedings of the ACM Multimedia 2010 International Conference* (New York, New York, USA, 2010), ACM Press, pp. 251–260.

[36] REIMERS, N., AND GUREVYCH, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference* (8 2019), 3982–3992.

[37] RÜGER, S. Multimedia information retrieval. *Synthesis Lectures on Information Concepts, Retrieval, and Services 1*, 1 (2009), 1–171.

[38] SCHROFF, F., KALENICHENKO, D., AND PHILBIN, J. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 815–823.

[39] SUN, C., MYERS, A., VONDRICK, C., MURPHY, K., AND SCHMID, C. VideoBERT: A Joint Model for Video and Language Representation Learning. *Proceedings of the IEEE International Conference on Computer Vision 2019-Octob* (4 2019), 7463–7472.

[40] TAN, R., XU, H., SAENKO, K., AND PLUMMER, B. A. wman: Weakly-supervised moment alignment network for text-based video segment retrieval, 2019.

[41] VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, , AND POLOSUKHIN, I. Transformer: Attention is all you need. In *Advances in Neural Information Processing Systems* (2017), vol. 2017-Decem, pp. 5999–6009.

[42] XIE, S., SUN, C., HUANG, J., TU, Z., AND MURPHY, K. Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-offs in Video Classification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 11219 LNCS* (12 2017), 318–335.

[43] XU, J., MEI, T., YAO, T., AND RUI, Y. Msr-vtt: A large video description dataset for bridging video and language. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR).

[44] YAN, C., TU, Y., WANG, X., ZHANG, Y., HAO, X., ZHANG, Y., AND DAI, Q. STAT: Spatial-Temporal Attention Mechanism for Video Captioning. *IEEE Transactions on Multimedia 22*, 1 (1 2020), 229–241.

[45] ZHANG, L., NIZAMPATNAM, S., GANGOPADHYAY, A., AND CONDE, M. V. Multi-attention Networks for Temporal Localization of Video-level Labels.

Curriculum Vitae

Jesús Andrés Portillo Quintero was born in the hot lands of Cd. Obregón, Sonora, México on January 7th of 1996. He earned the Industrial and Systems Engineering degree from the Instituto Tecnológico y de Estudios Superiores de Monterrey, Guadalajara Campus in December 2018. He was accepted in the graduate program in Computer Sciences in August 2019.

This document was typed in using LaTeX 2_ε[1] by Jesús Andrés Portillo Quintero.

[1]The template `MCCi-DCC-Thesis.cls` used to set up this document was prepared by the Research Group with Strategic Focus in Intelligent Systems of Tecnológico de Monterrey, Monterrey Campus.