Spatial-Temporal Reasoning in Symbolic Neural Network for Semantic Interpretation of Videos

UFAM

Milena Rodrigues Tenorio

Supervisor: Prof. Dr. Edjard Souza Mota

Institute of Computing
Federal University of Amazonas

This Thesis is submitted for the degree of
Master in Computing

Manaus-AM June 2022

MILENA RODRIGUES TENORIO

Spatial-Temporal Reasoning in Symbolic Neural Network for Semantic Interpretation of Videos

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Informática do Instituto de Computação da Universidade Federal do Amazonas, como requisito parcial para a obtenção do título de Mestre em Informática.

Master's thesis submitted to the Graduate Program in Informatics of the Computer Science Institute of the Federal University of Amazonas, as partial requirement for obtaining the title of Master in Informatics.

Supervisor: Prof. Dr. Edjard Souza Mota

Manaus / AM
2022

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Tenorio, Milena Rodrigues
T312s   Spatial-temporal reasoning in symbolic neural network for semantic interpretation of videos / Milena Rodrigues Tenorio . 2022
67 f.: il. color; 31 cm.

Orientador: Edjard de Souza Mota
Dissertação (Mestrado em Informática) - Universidade Federal do Amazonas.

1. Neural Symbolic. 2. Semantic interpretation of video. 3. Spatial-temporal reasoning. 4. Neural network. I. Mota, Edjard de Souza. II. Universidade Federal do Amazonas III. Título

PODER EXECUTIVO
MINISTÉRIO DA EDUCAÇÃO
INSTITUTO DE COMPUTAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

UFAM

FOLHA DE APROVAÇÃO

"Spatial-Temporal Reasoning in Symbolic Neural Network for Semantic Interpretation of Videos"

MILENA RODRIGUES TENÓRIO

Dissertação de Mestrado defendida e aprovada pela banca examinadora constituída pelos Professores:

Prof. Edjard de Souza Mota - PRESIDENTE

Paulo César da Rocha Fonseca

Prof. Paulo Cesar Fonseca - MEMBRO EXTERNO

Prof. Rafael Giusti - MEMBRO INTERNO

Manaus, 18 de Março de 2022

I would like to dedicate this thesis to my God, my loving parents and brother.

I would like to express my gratitude to my thesis advisor prof. Edjard Mota for all help and patience, as well as to UFAM and IComp for the opportunity, and CAPES for the 88882.461676/2019-01 scholarship.

I would like to thank the committee, profs Paulo Cesar Fonseca and Rafael Giusti for their inputs, which helped me to improve this work.

Finally, thank you to my family for their love and support.

Resumo

O campo de estudo sobre Interpretação Semântica de Vídeos procura maneiras de modelar as informações existentes nos vídeos. Os métodos existentes podem ser divididos em métodos genéricos e especializados, o primeiro é capaz de categorizar as informações com eficiência e o especialista não tem um bom desempenho para dados genéricos. Uma maneira de os pesquisadores lidarem com esse impasse, em outros campos de estudo, é usar o conhecimento e as restrições sobre ele. Para isso, usamos o raciocínio neural-simbólico. Nossa hipótese é usar uma rede neural simbólica para extrair informações de imagens de um vídeo para modelar essas informações, e enfim realizar raciocínio para extração da descrição semântica. Para tal propósito foram escolhidos três principais etapas (1) identificação dos objetos nas imagens do vídeo, (2) identificação das relações espaciais em grupos de frames e (3) analise das relações temporais encontradas, através dessas etapas identificamos com esta pesquisa que é possível inferirmos as ações que acontecem em um vídeo através do algoritmo proposto.

Palavras-chave: Neural-simbolico, Interpretação Semântica de Vídeo, Raciocínio Espaço-Temporal.

Abstract

The Semantic Video Interpretation field of study looks for ways to model the information in videos. Existing methods can be divided into generic and specialized methods; the former can efficiently categorize information while the latter does not perform well for generic data. One way for researchers to deal with this impasse, in other fields of study, is to use the knowledge and basic restrictions on it. For this, we use neural-symbolic reasoning. Our hypothesis is to use a neural-symbolic network to extract information from images in a video to model this information, and finally perform reasoning to extract the semantic description. For this purpose, three main steps were chosen: (1) identification of the objects in the video images, (2) identification of the spatial relations in frame groups, and (3) analysis of the temporal relations found.

Keywords: Neural-symbolic, Semantic Interpretation of Video, Spatial-Temporal Reasoning.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Video semantic interpretation aims to recognize what happens in a video. This involves recognizing the objects in the video and the actions that occur in it. Recognition of the actions of what happens in a video and the semantic understanding of what happens in the objects present are important fields of study of intelligent video analysis.

One of the challenges identified is related to the recognition of human behavior due to the low recognition efficiency and low accuracy of some proposed algorithms that will be discussed in this dissertation. Video semantic interpretation is important for several reasons because the studies for intelligent video analysis describe videos in different contexts with different goals.

This study specialization aims to extract information from a video, allowing it to be described according to the objective under study during this extraction. This can occur through the analysis of the objects identified in the video and the actions they perform in the captured time.

As an example, consider a hypothetical situation of a video showing traffic on an avenue. A company responsible for enforcing traffic laws would interpret if any laws were broken in traffic. However the companies that suggest travel routes for drivers would interpret the same data in order to identify whether traffic is intense on this avenue. The goals of an interpretation of data can be differentiated by contexts as in a hypothetical situation of a video showing traffic on an avenue. A company responsible for enforcing traffic laws would interpret if any laws were broken in traffic. However the companies that suggest travel routes for drivers would interpret the same data in order to identify whether traffic is intense on this avenue.

Through these examples we identified that the way of studying the data is different depending on how the data is being analyzed. This means that the descriptions obtained were different. But if someone questioned how the result of such an analysis was obtained, we would need to explain the method of study. Because the description is the result of the analysis made by the algorithm and the explainability is the human understanding of how this result was obtained.

As we have the objective of *description and explainability* in this research, we opted for the semantic interpretation of the video. Because we want to obtain an algorithm capable of describing the events of a video in such a way that it is possible to understand how the result of the description was obtained. In other words, description is about what we want to interpret in the video semantically by recognizing the objects present in the video and describing the actions they perform. Explainability, in turn, is how comprehensive an algorithm answer is easily understood by people.

This interpretation must have information that allows the understanding of what happened in the video and what relations occurred between the objects during the identified action.

The importance of semantic extraction with explainability occurs through the important possible applications for such a method. The development of explainable methods aims to explain to a person how the process of the algorithm in question occurred.

In this way information that demonstrates how our method arrived at an identification makes it explainable at some level. As soon as we developed this method initially for a small scope, with the possibility of being expanded to several other applications, we would bring the explanation for the identification of actions in videos.

Obtaining a video semantic description is a complex task because the data information used involves different types of data. The data information used for this field of study includes image analysis, study of temporal and spatial relations, data ontology, object identification, video analysis and among others.

The semantic description of videos will be more convincing if explainability could support decisions made by the algorithm. The semantics is focused on the description of meanings. Our purpose is to identify the actions in a video through the analysis of the movements that occur in it so that the meaning of these movements justify the identification of the action.

Because of this we identified that the explainability applied in our method would bring the understanding of how the actions were identified.The expected explainability for this method is the understanding of how a person can understand how the algorithm completes its reasoning about the actions identified by the model.

Using the proposed method, we can apply different types of scopes in videos and in addition to all kinds of relationships that can develop over time. Identify the behavior of people in a bank with the intention of robbery, in videos that analyze over time the face of a driver if he presents tiredness or sleep that causes danger of accident, evolution of CT scans of cancer patients over the years, analysis of voters' tweets during the electoral campaign when they have contact with fake news.

In this way, we chose to analyze short videos of soccer contexts. We want to specifically identify actions between pairs of objects. The actions between the player and the ball would be: losing the ball, controlling the ball and passing the ball. Actions between two opposing players would be: approaching the opponent, attacking the opponent and sounding out the opponent.

Although video research is old, one of the first successful descriptive methods are SVOs. SVOs are methods based on tuples (Subject, Object, Verb) used specifically for video description for Rohrbach [6] in survey Aafaq [7]. Among them, we can highlight the work of Koller [8], who developed a system that was able to characterize vehicle movement in real traffic scenes using natural-language verbs, and Brand [9], which described a series of actions in semantic tag summaries to extract the description of actions.

Current solutions, as in Cohn [10], perform the identification of objects in a video frame, prioritizing the identification of the positioning of elements in an image in a qualitative way, and after that infer the sequence of actions that can happen with objects from that video.

Due to these considerations, our proposal to bring the semantic description of soccer videos is based on a *neuro-symbolic network*. Most deep neural networks are considered black boxes, which means that their output is difficult for humans to interpret. In contrast, logical expressions are considered more understandable, as they use symbols that are semantically close to natural language rather than distributed representations, [11]. Thus we find the starting point for the neuro-symbolic approach.

To get a broader view, but within our context of division and conquest, we looked for an area of knowledge that has a lot of influence for our purpose, we concluded that for video analysis we would be dealing with the field of study called Semantic Interpretation of Images (SII). This is the study area that seeks to extract the meaning of information from visual data, generally extracting semantic descriptions of images by modeling, allowing the use of images for various applications.

Methods that use the SII approach have the fundamentals of Artificial Intelligence (AI) and Machine Learning (ML) called Statistical Relational Learning (SRL), as described in Donadello [12] which are approaches that deal with uncertainty domain models and complex relational structures. This was extended in [2], through the use of background knowledge in their framework. This classification approach is called a Logic Tensor Network (LTN).

In Donadello [13] it is shown that SII can be yielded from a labelled graph, an image or scene. This graph can be interpreted as the semantic description of the image content.

This graph is a direct correspondence between the low-level features of an image and its high-level semantic description.

However this approach is only for images, this means it is only used in static contexts. LTN is not yet able to compute video semantic recognition because it does not have a temporal dimension of videos. As LTN is a neural-symbolic approach, it allows inductive learning and reasoning. Through the information learned and the possibility of its application in explainable AI methods, this approach relates the computationally processed context with the external resources of available prior knowledge, see Tiddi [14]. This makes this approach closer to building human thinking than other machine learning models such as the end-to-end approach.

The challenge of bringing computational reasoning closer to the human for knowledge of events began in the Tenorio research [15]. The initial discussion addressed the topic of temporal symbolic reasoning. Through a closed scenario, it was possible to evidence an explosion of hypotheses, at the same time that logical rules were extracted even with a high level of expressiveness.

Such an experiment was important for understanding how the symbolic model of spatial-temporal information behaves in a purely symbolic approach. Allowing us to understand that when dealing with two-dimensional information, in this case, such as space and time, it is necessary to have a model capable of relating these dimensions in a coherent way. However, the advances achieved by AI and ML demonstrate a high impact on the challenges encountered in the Tenorio research [15]. In addition to the growing concern about reliability, security, and interpretability of the results obtained from AI methods.

Therefore, there is a need for a better representation of knowledge and reasoning to integrate with AI deep learning and explainability. Neural-symbolic computing has been an active area of research for many years. It seeks robust neural network learning with reasoning and explainability through symbolic representations for neural network models, discussed in Garcez [16].

Considering all this, the main objective is to develop an algorithm with better semantic description of video, being able to identify objects, relations between them and infer their respective actions.

As specific objectives, we intend to: develop an algorithm with more descriptive and interpretive video information for humans, being able to identify objects, relations between them and infer the respective action. Our specific goals aim to identify the objects and their positions in the video, using a framework that is superior to others for this purpose; provide the identification of spatial relations in symbolic neuro-networks; adapt LTN relations based on the QQSTR method; model the identified information in a data structure that allows the identification of the action that occurred in the video.

The remainder of this work is organized as follows: Chapter 2 presents Fundamental Concepts necessary for understanding the problem, and the methods used to solve it; in chapter 3 we describe the proposed model, in addition to demonstrating its application in a toy generic example for, finally, its application in the context of our final objective, actions in soccer; the next chapter, Experiment, describes how knowledge construction process, the associated logical rules, and the construction of the proposed model for the context of actions in soccer, which was the chosen context; ending in chapter 5, where we present the conclusions of our research.

Chapter 2

Summary of Related Works

Given the information about the purpose and challenges to be faced in the elaboration of this method, below we will briefly describe the main base works for this research and their contributions to it. The main works are LTN for SII by Donadello [2], QQSTR by Cohn [4] and Activity Graph by Sridhar [17].

The Logic Tensor Network (LTN) is a framework that uses neural symbolism, it has been shown to be superior to some methods, as in the study by Donadello to identify the *part-of* spatial relationship between objects in images. Qualitative and Quantitative Spatial-Temporal Relations (QQSTR) is an approach that models spatial-temporal information using qualitative and quantitative characteristics. And finally, Activity Graphs that model spatial relationships in time using the concept of directed graphs.

In addition to these methods, we highlight the importance and contribution of the YOLO framework [3]. Through it we will identify the objects in the videos and classify them. In addition, it will play a fundamental role in the execution of our method to relate the different approaches that we will use.

As described, LTN is used only in static contexts and has already been applied to images, but videos have a temporal dimension. This dimension is addressed by the two other fundamental works QQSTR and Activity Graphs. And we will use YOLO to adapt this dimension to be used by LTN. In the following chapters we will further discuss this need and implementation.

The relevance of LTN use is reinforced in Donadello's research [2]. Using the neural-symbolic approach of the LTN structure, he shows that it is possible to identify spatial relationships between pairs of previously identified objects in a way superior to methods such as Fast-RCNN.

For such relevance, we adopted the following adaptation to list the works used: YOLO, in addition to identifying and categorizing the objects of the videos, will be able to fragment the video into frames, that is, images that are capable of being processed by the LTN.

The LTN in turn will be able to reason about the relations between pairs of video objects. These spatial relationships will be based on the approach used by Cohn, so that they are qualitative relationships that allow explainability to continue in our approach in a judicious way.

Finally, the temporal relations received by YOLO and the spatial relations reasoned by the LTN will be structured in Activity Graph.

What we propose is an extension of the LTN's Semantic Image Interpretation to apply it in the context of spatial-temporal reasoning. The justification for this choice lies in the better learning results of LTN when compared to other semantic image description frameworks.

The importance of using a framework with a symbolic neural network lies in the strong representation that this approach has. Allowing that the neural network of this framework is not a total black box, making the inferences reasoned by our hypothesis a little more understandable to the human being than the other models.

We intend to demonstrate an algorithm for semantic analysis of soccer videos, such an algorithm will have several steps to reach the final objective. We will have the pre-processing of the video through the YOLO framework, to then be analyzed by the LTN that will reason by spatial relations based on the QQSTR approach. Finally, the spatial relationships will be structured in an Activity Graph with the equivalent temporal relationships to infer the actions of soccer videos.

Chapter 3

Fundamental Concepts

The extraction of information from images and videos formed by raw data (without metadata) can be related to the symbolic approach to obtain the level of explainability and reasoning as described in Donadello [18].

There are several advantages of extracting information from symbolic knowledge as a possible advancement towards explainable AI, making such an approach relevant in neural-symbolic environments, as described by Garcez [19]. This possibility of extracting symbolic knowledge from pure data in artificial neural networks provides the investigation of how to improve the explainability of these methods.

In this way, we approach in this chapter, the concepts that underlie the relevance of neural-symbolic AI, the LTN framework, and its applications in the semantic analysis of images. We discuss other related works such as YOLO, space-time reasoning, and the semantic interpretation of the video.

3.1 Artificial Neural Networks

The human brain processes information in a completely different way than conventional digital computers. It is in the interest of Artificial Intelligence to propose requirement specifications and express mental models of reasoning.

Starting from that point, Arrieta [20] addresses information about the consensus that has been created on the importance of explainability. That is, intelligent machines are endowed with learning, reasoning, and adaptability so that it is still possible to explain to human beings how these factors are computed.

Although the first systems provided with artificial intelligence were easily interpretable, in recent years black box systems such Artificial Neural Network, also called the Connectionist System, are a powerful approach to Machine Learning, inspired by biology and neurology as demonstrated in Bader [21]. Black box systems are those that do not allow the human interpretability of their results, which, unlike white-box systems, are easily interpretable.

Among the black box systems, Deep Learning (DL) models emerged through the union of efficient learning algorithms and the ability to process several and numerous parameters. Such conditions have made DL applications, such as Deep Neural Networks (DNNs), considered complex black-box models explain Arrieta [20].

Among the AI methods, we have Connectionist AI, inspired by intelligent behavior models. It structures parallel information strongly connected to each other just as the brain acts between its neurons. The inspiration of this structure is so directly related to the human nervous system that its components are:

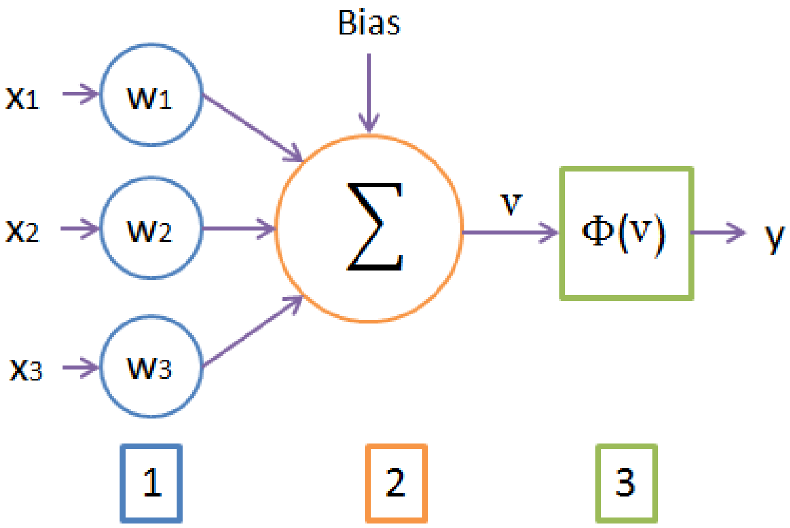


Figure 3.1: Illustration of the behavior of a neuron, adapted from [1]

Besides this method, we also emphasize that Neural Networks are very well distributed parallel processors, formed by simple processing units (artificial neurons) that are capable of storing knowledge and making it available for use. As illustrated in Figure 3.1, the behavior of a neuron can be analyzed in three different steps described by Borges [1]:

1. The synapse (connection) represents the input of a neuron. It is defined by a weight, which is a real number that will multiply the input values of that synapse;

2. Followed by a combination of weighted inputs and a bias. These values are usually added to get a new value that will go to the next step v;

3. Finally, an activation function ϕ is applied to the value v in the previous step. This is how the neuron output value called neuron activation is defined.

The comparison between neural networks and the human brain can be done in two main learning aspects: learning from the environment and storing acquired knowledge. This knowledge is obtained by connecting neurons, also called *synaptic weights*.

Its main advantages are its strong parallelism, fault tolerance due to its robust learning, efficiency in inductive learning, and generalization capability. According to Garcez [22] they have been used in a variety of tasks, including pattern recognition, robot control, DNA sequence analysis, and time series analysis and prediction.

Among the main characteristics of artificial neural networks, we can accentuate the natural tendency to empirical learning, through the use of learning algorithms. This learning generally occurs by adapting the synaptic weights and the values of the neurons according to the output error, as explained by Borges [1].

Therefore, we will highlight some attributes that are important for deep learning, Battaglia [23] define as entities, relations, and rules. An entity is an element with attributes, a relation is a property between entities, and a rule is a function that maps entities and/or relations to each other.

3.1.1 Types of Artificial Neural Networks

We initially identified that methods before Artificial Neural Networks valued representation due to the computational power of the time. Currently, as more robust technologies in terms of processing time and information storage capacity, deep learning methods have emerged that seek to be end-to-end methods, that is, they do not emphasize the representation of information, but the computation that is possible to perform on the data.

For this reason, this subsection explains the different purposes that neural networks can have, the fields of study in this area, and the use of different connectionist architectures. Battaglia [23] cites some examples as:

- *Fully connected layers*: each neuron has an all-to-all relation and the rules are defined by weights and bias, as illustrated in Figure 3.2;

- *Covolutive layers*: entities are units like pixels of an image, but with sparse relations and reflecting relational rules in neurons according to their locations, that is, the closer the neurons, the greater the influences between them. See 3.3;

- *Recurring layers*: a sequence of steps is implemented, with inputs and hidden states in each of them. In order to combine so many different aspects, the input entities are

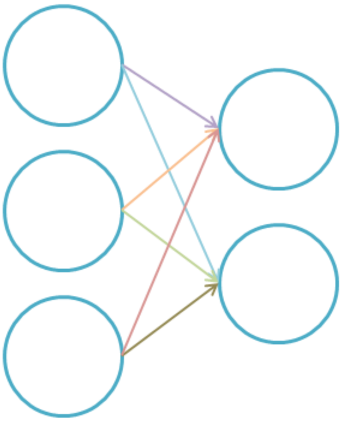


Figure 3.2: Example of a neural network's fully connected layers

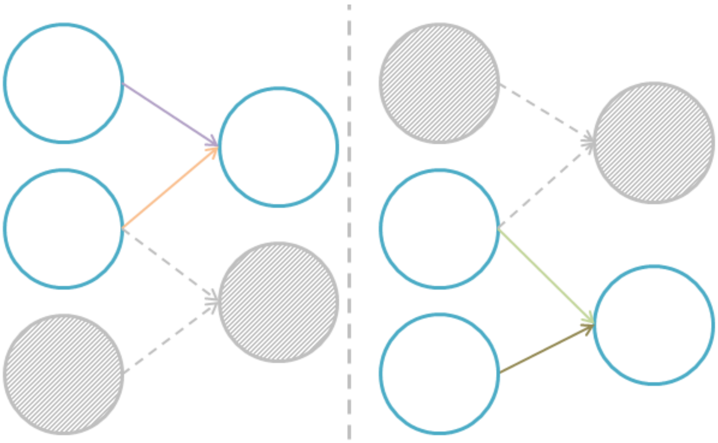


Figure 3.3: Example of a neural network's covolutive layers

connected to the output entities from the previous step. That way the hidden states can be updated in the next state as they will be used as input, see 3.4.

Still addressing the concepts of neural networks, there is the importance of the black box behavior of this approach, that is, without explanation for humans. But it is possible to recognize that such an AI is capable of recognizing patterns and reasoning about data, and is endowed with learning.

Among other characteristics of neural networks, according to Battaglia [23], the principle of combinatorial generalization shows the construction of new inferences, predictions, and behaviors from known blocks. However, for humans, Cohen [24] explains this, which relies on the cognition engine to represent structures and reason about existing relations.

Therefore, it is evident that Neural Networks are methods of computational learning that are inspired by the human brain in order to improve reasoning and learning through computer

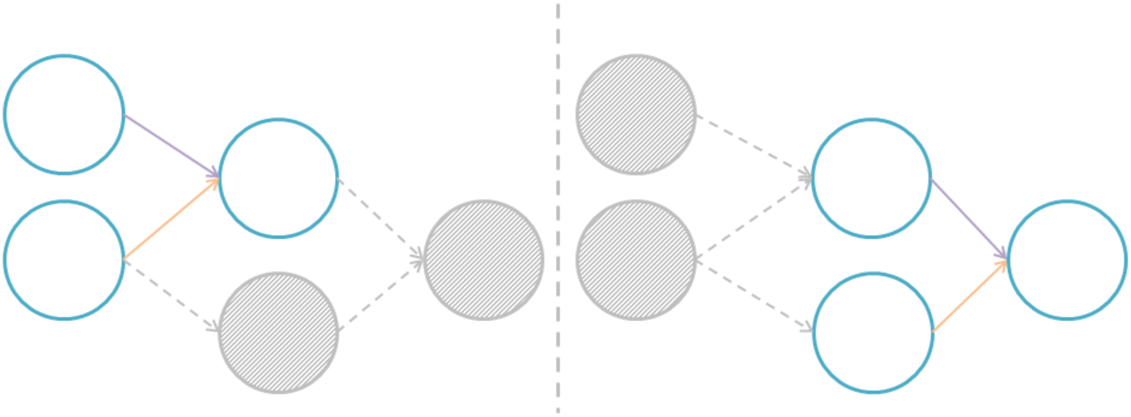


Figure 3.4: Example of neural network recurring layers

algorithms, aiming at the expected result without prioritizing the explainability of how it was obtained.

Next, we present subsections that discuss a specific category of artificial neural networks, the Symbolic Neural Network; and the framework that we propose to use in this research, the Logic Tensor Network, which is a framework with a symbolic neural network.

3.2 NeSy AI

According to Sarker [25], Neuro-Symbolic Artificial Intelligence (NeSy AI) is a subfield of the field of Artificial Intelligence. NeSy AI can be defined as the association of symbolic methods and methods based on artificial neural networks. Sarker cites:

> "The term neural in this case refers to the use of artificial neural networks, or connectionist systems, in the widest sense. The term symbolic refers to AI approaches that are based on explicit symbol manipulation. This in general would include things like term rewriting, graph algorithms, and natural language question answering. It is often more narrowly understood, though, as a reference to methods based on formal logic, as utilized, for instance, in the subfield of AI called Knowledge Representation and Reasoning. The lines easily blur, though, and for the purposes of this overview, we will not restrict ourselves to logic-based methods only."

One of the main differences between approaches is the representation of information within an AI method, and it is precisely this difference that causes the overlap of one NeSy approach with the others.

In symbolic systems, the representation of information is totally direct, it is interpretable by a human being, in addition, it can be manipulated and logically inferred.

However, neural systems have a strongly connected representation between neurons and their simultaneous activation between them, thus what we call a black box, that is, difficult to understand for a human being, as discussed in section 3.1.

3.3 Logic Tensor Network

LTN is a framework that integrates Artificial Neural Networks, SRL with First Order Fuzzy Logic. Through constraints and logical formulas reasoning about the properties of information, it is possible to get efficient learning from noisy data described in Serafini [12].

This is a framework capable of identifying binary relations between constants, through input data properties, prior knowledge, and logical descriptions, as described by Donadello [2]. Donadello [18] carried out a study on such capabilities in a static context with one-dimensional approaches, that is, studies with pure data.

The advantage of LTN over other frameworks comes from the combination of visual data entry with Background Knowledge (BK) corresponding to the data. This allows the framework to avoid the "zero-shot" learning that occurs when a framework learns without any prior information to guide its reasoning, explain Donadello [18].

With an LTN, it is possible to have a relation between logical rules and the learning that takes place on the network, Making it's black-box grayer, that is, with more explainability than common neural network models. This approach of using the learning approach based on logical rules characterizes LTN as a neural-symbolic approach.

According to Bennetot [26] the neural-symbolic approach is important because of (1) the possibility of logical induction for reasoning about the information learned, (2) the possibility of its application in eXplainable AI methods, (3) the extraction of logical rules, and (4) the modeling of information closer to human thought.

In Donadello [2] the LTN application is described to combine visual and symbolic knowledge in the form of logical axioms to solve two problems of SII: (1) the classification of objects identified in the images, and (2) identifying *part-of* relations between the objects that compose it.

Thus, an LTN aims to compensate the mismatch between low-level (numeric) characteristics that can be observed in an image and the high-level semantic descriptions associated with objects present in it, this is the semantic gap, through the use of background knowledge according Donadello [2].

3.3.1 Definitions of LTN

The LTN syntax is the same as the First-Order Predicate Language (PL): definition is comprised by the terms $\{C,F,P\}$, where C are constant of symbols; F are functions of real numbers; P are predicates interpreted as functions on real vectors of [0, 1] for relations, whose the *domain* is a subset of R^n.

Because it is a Predicate Language, LTN also needs to satisfy the same conditions presented by Donadello [13]:

- $\mathscr{G}(c) \in \mathbb{R}^n$ for every constant symbol $c \in \mathscr{C}$;
- $\mathscr{G}(f) \in \mathbb{R}^{n.a(f)} \rightarrow \mathbb{R}^n$ for every functional symbol $f \in \mathscr{F}$;
- $\mathscr{G}(P) \in \mathbb{R}^{n.a(P)} \rightarrow [0,1]$ for every predicate symbol $P \in \mathscr{P}$

Definition 1 (N-grounding) *A formula is interpreted by its degree of veracity. LTN uses the term* grounding *as a synonym for logical interpretation in a "real world", $\mathscr{G}$ (with $n \in \mathbb{N}$, $n > 0$) captures the correlation between objects, their category properties and relation to these conditions:*

1. *G(c)$\in \mathscr{R}$, for every* c $\in$ C
2. *G(P) $\in \mathscr{R}^{n.\alpha(P)} \rightarrow [0,1]$, for every $P \in \mathscr{P}$*

Given a grounding $\mathscr{G}$ and let terms*($\mathscr{P}$ $\mathscr{L}$) = t_1, t_2, t_3 ..., the induction is defined as follows:*

$$\mathscr{G}(P(t_1,...,t_m)) = \mathscr{G}(P)(\mathscr{G}(t_1),...,\mathscr{G}(t_m))$$

(Equation 1)

Under such conditions LTN has the exact grounding of an unknown symbol ϕ, but it is known that it can be obtained by finding a set of real-valued parameters, that is, via learning.

To emphasize this fact, LTN adopts the notation by Badreddine [27]. This process requires a t-norms operator used in Fuzzy Logic. In this case, Donadello [13] describes using Lukasiewicz t-norm for the semantics of non-atomic formulas.

Due to such grounding, LTN defines Lukasiewicz t-norm according to the following functions:

$$\mathscr{G}(\phi \rightarrow \psi) = min(1, 1 - \mathscr{G}(\phi) + \mathscr{G}(\psi))$$
$$\mathscr{G}(\phi \wedge \psi) = max(0, \mathscr{G}(\phi) + \mathscr{G}(\psi) - 1)$$
$$\mathscr{G}(\phi \vee \psi) = min(1, \mathscr{G}(\phi) + \mathscr{G}(\psi))$$
$$\mathscr{G}(\neg\phi) = 1 - \mathscr{G}(\phi)$$

(Equation 2)

However, LTN is not limited to such t-norms; there are also functions for the interpretation of quantifiers, for example, the interpretation of $\forall$ is:

$$\mathscr{G}(\forall x\phi(x)) = inf\mathscr{G}(\phi()t)|t \in term(\mathscr{P}\mathscr{L})$$

(Equation 3)

The definitions mentioned above do not tolerate exceptions; however, LTN is able to handle this type of information. Because this framework can handle exceptions by assigning higher truth outliers to the formula $\forall x\phi(x)$, providing learning with the required exception, if many examples satisfy: $\phi(x)$. See:

Definition 2 *Let* mean$_p$*($x_1,...,x_d$)* = $\left(\frac{1}{d}\sum_{i=1}^{d} x_i^p\right)^{\frac{1}{p}}$*, with* $p^1 \in \mathscr{Z}$, d $\in \mathscr{N}$*, the grounding for* $\forall \mathrm{x}\phi(\mathrm{x})$ *is:*

$$\mathscr{G}(\forall \mathrm{x}\phi((x))) = \lim_{d \to |term(\mathscr{P}\mathscr{L})|} mean_p(\mathscr{G}(\phi(t_1)),...,\mathscr{G}(\phi(t_d)))$$

(Equation 4)

The grounding of a quantified formula $\forall x\phi(x)$ is the mean of the d groundings of the quantifier-free formula $\phi(x)$. Regularity is necessary for a suitable function for grounding. Let $b \in C$ refer to a bounding box constant containing a dog as illustrated in 3.5. Let v = $\mathscr{G}(b)$ be its feature vector, then it holds that $\mathscr{G}$(Dog)(v) $\approx$ 1.

Moreover, for every bounding box with feature vector v' similar to v, $\mathscr{G}$(Dog)(v') $\approx$ 1 holds. These functions are learned from data by tweaking their inner parameters in a training process.

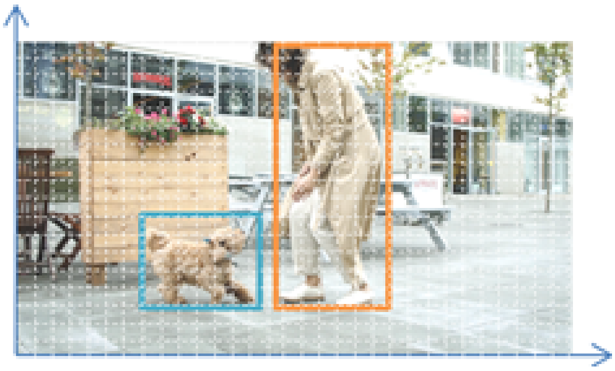


Figure 3.5: Example of bounding box notation in a video frame

In addition to the functions, we have the *predicate symbols*, these are effective architectures for relational learning. Let b_1,...,$b_m \in \mathscr{C}$ with feature vectors $v_i = \mathscr{G}(b_i) \in \mathscr{R}^n$, with i = 1,...,m, and v = <v_1;...;v_m >is a m-ary vector given by the vertical stacking of each vector v_i.

In this way, there is also the grounding associated with the predicates. The grounding $\mathscr{G}$(P) of an m-ary predicate $\mathscr{P}(b_1,...,b_m)$ where σ is the sigmoid function:

$$\mathscr{G}(P)(v) = \sigma\left(u_P^T tanh\left(v^T \mathscr{W}_P^{[1:k]} v + V_P v + b_P\right)\right)$$
(Equation 5)

Finally, for the predicate, we have the following parameters: $\mathscr{U}_P \in \mathscr{R}^k$, a 3-D tensor $\mathscr{W}_P^{[1:k]} \in \mathscr{R}^{k\times mn\times mn}$, $\mathscr{V}_P \in \mathscr{R}^{k\times mn}$ and $b_P \in \mathscr{R}^k$. The parameter $\mathscr{U}_P$ computes a linear combination of the quadratic features returned by the tensor product. With Equations (1) and (5) the grounding of the closed terms are computed, and the atomic formulas are combined using a specific operator t-norm, see Equation (2).

These are the ways to build the formulas to be used in LTN. Through this, symbolic neural networks can be built. In the next section, LTN learning will be discussed.

3.3.2 LTN Learning

LTN learning occurs through the construction of groundings, a process which involves optimizing the truth values of the formulas discussed in the previous section. Through this, an LTN knowledge base is built, also known as *grounded theory*.

Through the divide and conquer approach, a *partial grounding* $\hat{\mathscr{G}}$ is a grounding defined on a subset of the signature of $\mathscr{PL}$, to finally unite in grounding $\mathscr{G}$ for $\mathscr{PL}$. Donadello [13] define this grounding as a *completion* of $\hat{\mathscr{G}}$ (in symbols $\hat{\mathscr{G}} \subseteq \mathscr{G}$) if $\mathscr{G}$ coincides with $\hat{\mathscr{G}}$ on the symbols where $\hat{\mathscr{G}}$. See the definitions about this in LTN:

Definition 3 *A* grounded theory GT *is a pair* $<\mathscr{K},\hat{\mathscr{G}}>$*with* $\mathscr{K}$ *a set of closed formulas and* $\hat{\mathscr{G}}$ *a partial grounding.*

Definition 4 *A grounding* $\mathscr{G}$ *satisfies a grounded theory* $<\mathscr{K},\hat{\mathscr{G}}>$*if* $\hat{\mathscr{G}} \subseteq \mathscr{G}$ *and* $\mathscr{G}(\phi) = 1$, *for all* $\phi \in \mathscr{K}$. *A grounded theory* $<\mathscr{K},\hat{\mathscr{G}}>$*is* satisfiable *if there exists a grounding* $\mathscr{G}$ *that satisfies* $<\mathscr{K},\hat{\mathscr{G}}>$.

Grounding Theory (GT) $< K, \hat{G} >$ is defined as an extension of partial grounding $\hat{G} >$ on all possible grounds. So that all instances of clauses in K are satisfied within the range, making grounding satisfactory.

But this is not practical, to verify the satisfiability in the grounding, a correlation must be captured between the quantitative attributes of an object and its relational properties. Serafini [28] explains that to limit the number of instances of clauses, which can be infinite, we generally consider the instances of each clause to a certain depth.

According to the above definition, the satisfiability of $<\mathscr{K},\hat{\mathscr{G}}>$can be obtained by searching for a grounding $\mathscr{G}$. That extends $\hat{\mathscr{G}}$ such that *every* formula in $\mathscr{K}$ has value 1 when a grounded theory is not the GT.

Definition 5 *Let $<\mathscr{K},\hat{\mathscr{G}}>$be a grounded theory. This definition is related to the problem of the* best satisfiability*, this amounts to searching an extension $\mathscr{G}^*$ of $\hat{\mathscr{G}} \in G$ (the set of all possible groundings). Serafini [28] explains that, maximizing the satisfiability error on the set, and minimizing the truth value of the conjunction of the formulas in $\mathscr{K}$:*

$$\mathscr{G}^* = \underset{\hat{\mathscr{G}}\subseteq\mathscr{G}\in G}{\arg\max}\,\mathscr{G}\left(\begin{matrix}\phi \in \mathscr{K}\,\phi \\ \wedge\end{matrix}\right) \textit{(6)}$$

The *maximum satisfiability problem* is an optimization problem on the set of parameters to be learned. Let $\Theta = \mathscr{W}_{\mathscr{P}}, \mathscr{V}_{\mathscr{P}}, b_{\mathscr{P}}, u_{\mathscr{P}} | \mathscr{P} \in \mathscr{P}$ be the set of parameters. Let $\mathscr{G}(\cdot|\Theta)$ be the grounding obtained by setting the parameters of the grounding functions to Θ.

The best satisfiability problem tries to find the best set of parameters Θ with $\lambda \parallel \Theta \parallel_2^2$ a regularization term:

$$\Theta * = argmax_{\Theta} G\left(\bigwedge\nolimits_{\phi\in\mathscr{K}} \phi|\Theta\right) - \lambda \parallel \Theta \parallel_2^2 \text{ (7)}$$

Through such learning, it is possible to identify different scopes of application of the use of LTNs as mentioned initially in this section. Next, the application of LTN in the specific scope of semantic image interpretation is explored.

3.3.3 LTN in Semantic Image Interpretation

With LTN, it is possible to perform linear regression, binary or multi-label classification, and relation learning. As our goal is to use LTN to analyze spatial relations between objects identified in a video, we focus on understanding the framework for relations between instances.

LTN learning for spatial relations in images is described by Serafini [29] using the following definitions :

Domain Two rectangles A and B represented by four real numbers and six spatial relations, as show in Figure 3.6:

Problem Given the above definitions it is necessary to provide examples of pairs for each relation and the background knowledge with the logical constraints, as shown in figure 3.6, making it possible to identify what is the relation between two new random objects.

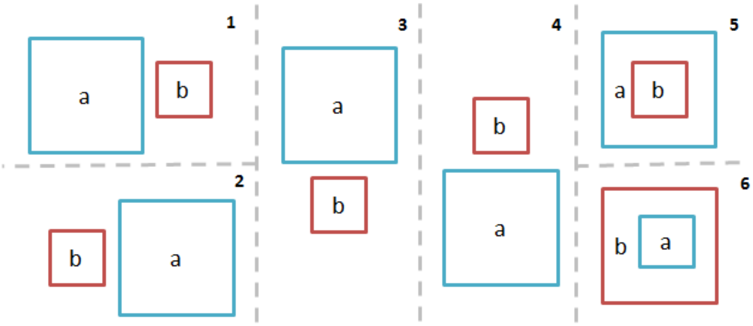


Figure 3.6: 1. a to the left of b, 2. a to the right of b, 3. a is a above b, 4. a is below b, 5. a contains b and 6. a is contained by b.

Language Defined as: $left(a,b)$, $right(a,b)$, $above(a,b)$, $below(a,b)$, $contains(a,b)$ and $in(a,b)$

Constraints Positive and negative examples of each spatial relations and axioms of spatial relations as: $\forall(x,y) : left(a,b) -> \ left(b,a)$

Donadello [2] uses LTN to describe the identification of the *Part Of* relation between pairs of objects. These objects are identified in the input image. Each object is associated with a bounding box and has its characteristics represented in an n-dimensional array of real numbers, such as:

$$<class^1(0,b1)1)...,class^n,(0,b1),x0,y0,x1,y1>$$

As relates to this form of object identification, we have the labels as semantic classes, or types of objects, identified in the bounding boxes. The terms $class^1(0,b1),...,class^n$ are the possible classes for the objects described by the background knowledge; (x0, y0) are the upper left and (x1, y1) are the lower right coordinates.

These characteristics are related to an ontology of the descriptive structure, which is presented by Serafini [29] and can be represented by a labeled directed graph, where the vertices are the bounding boxes, which represent each object identified in the initial image.

Such vertices are labeled with the type of object that exists in them; edges are the relations between pairs of objects, labeled by the reasoned binary relation. Each graph of this structure is called a *scene graph* [2], as show in Figure 3.7.

Given the information on how the LTN is composed and its behavior to reason about the data, we now emphasize that the language of the LTN undergoes some adaptations to be used in the research of Semantic Interpretation of Images. We emphasize them below:

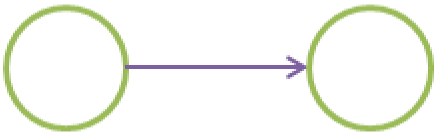


Figure 3.7: In this graph, the vertices represent the bounding box of each object in the image with their classification and the edges are labeled with the relations between pairs of objects

- C constants are all "bounding boxes" identified in the image, they are formed by semantic characteristics (possibility of belonging to the class of objects) and geometric (quadrant x_0, y_0, x_1, y_1)
- F empty, as the SII task of partial knowledge completion does not require function symbols;
- P the predicates are P_1, P_2, P1 is the set of unary predicates which are the object identification classes, as $P_1 = person, woman, child, animal, dog, cat$, and P2 is the set of binary predicates with the relations between pairs of bounding boxes, as $P_2 = hugging, jumping, running, talking$.

Semantic Interpretation of Images addresses the understanding of the relations *Part Of*. This relation is defined by checking the overlap between bounding boxes of objects in the image, and by analyzing the logical constraints that are provided as background knowledge. Because a dataset contains objects and pairs of objects labeled with a set of labels that are *relational data*, and *visual relation* are labels between pairs of bounding boxes that describe the semantic relations between the physical objects in the bounding boxes.

But when dealing with incomplete information, it is necessary for the LTN to be able to complete the missing information, this is called completion of the knowledge base. According Donadello [13] LTNs are developed to encode and solve this task with the help of logical constraint. This is done using grounded theory $\mathscr{T}_{SII} = \left\langle \mathscr{K}_{SII}, \hat{\mathscr{G}}_{SII} \right\rangle$, where $\mathscr{K}_{SII}$ is a LTNs knowledge base and $\hat{\mathscr{G}}_{SII}$ is a partial grounding. In order to complete a partial knowledge of the database, extend or land $\mathscr{G}^*_{SII}$ to $\hat{\mathscr{G}}_{SII}$, being:

$$\mathscr{G}^*_{SII}(C(b)) \to [0,1]$$
$$\hat{\mathscr{G}}_{SII}(R(b_1, b_2) \to [0,1]$$

3.3.4 Implementation of LTN for SII

To understand the implementation of LTN for SII, we first need to understand the state of the art on this type of study. Convolutional Neural Networks (CNNs) constitute the state-of-art models in all fundamental computer vision tasks, from image classification and object detection to instance segmentation as Arietta [20] describes. A CNN can be divided into two categories: 1) search to understand the decision process by mapping the output in the input space to see which parts of the input were used for the output; 2) find out how the network, and interpret how the middle layers, see the outside world.

These advantages over the use of the LTN framework for reasoning about images that we are addressing in this work were identified through experiments presented by Serafini [2]. In this experiment, they choose the "Part Of" relation to better handle the PASCAL-PART dataset, and ontologies were in WORDNET as cite Chen [30] and Fellbaum [31].

As relates to the "Part Of" relations, visual relations capture a wide variety of interactions between pairs of objects in images. This further emphasizes the importance of this field of study, because the better algorithms are developed for this study, the more scope applications can be explored, explains Lu [32]. With this, the Part Of can be used to represent many classes of relations between constants, including spatial relations.

However, it is worth noting that many other relations could have been included in this evaluation, but the time complexity of LTN grows linearly with the number of axioms. To train in a large visual set, many data labels can be affected by noise, such as missing or incorrect labels, non-localized objects, and disagreements between annotations. For example, labels for humans often confuse "Part Of" with the relation "Have". So to differentiate them would require a computational queue resource not interesting for our current research.

This LTN for the SII experiment used a PASCAL-PART dataset, the images of this dataset had bounding boxes for the objects, the categorization of the objects, and the part-of relations between pairs of bounding boxes. With these data, it was possible to reason the classification of the object type, and the detection of the "Part Of" relation. This was done through a set of bounding boxes detected by an object detector, in this case, Fast-RCNN.

In this way, each bounding box had the identification of its object type, and the "Part-Of" detection task occurred from a pair of bounding boxes, in case the object contained in the first is part of the object contained in the second.

LTN was used to solve both tasks: (1) because a type of bounding box and the part of the relation are not independent, (2) their dependencies are specified in LTN using background knowledge in the form of logical axioms. To show the effect of the logical axioms, two experiments with LTNs were carried out. The first had only examples of training types of objects and relation "Part Of" (T_{expl}). While in the second, logical axioms were added about types and "Part Of" (T_{prior}).

According to Donadello [2], the LTNs were set up with $TensorFlow^{TM}$ with layers as $k = 6$ and a regularization parameter $\lambda = 10^{-10}$. They choose Lukasiewicz's T-norm $(\mu(a,b) =$

$max(0, a+b-1))$ and use the harmonic mean as aggregation operator. The experiment ran 1000 training epochs of the RMSProp learning algorithm available in $TensorFlow^{TM}$. The results obtained were compared with those made in the Fast RCNN framework to classify the type of object and the "inclusion radius" (ir) to detect "Part Of". By the following definition: if the ir is greater than a certain limit th (in this case, $th = 0.7$ was adopted), the bounding boxes will have the relation "Part Of".

Every bounding box "b" is classified into $\mathscr{C} \in \mathscr{P}_1$ if $\mathscr{G}(\mathscr{C}(b)) \geq th$. With this, a bounding box can be classified into more than one class. For each class, precision and recall are calculated in the usual way. This allows a bounding box to have several classes.

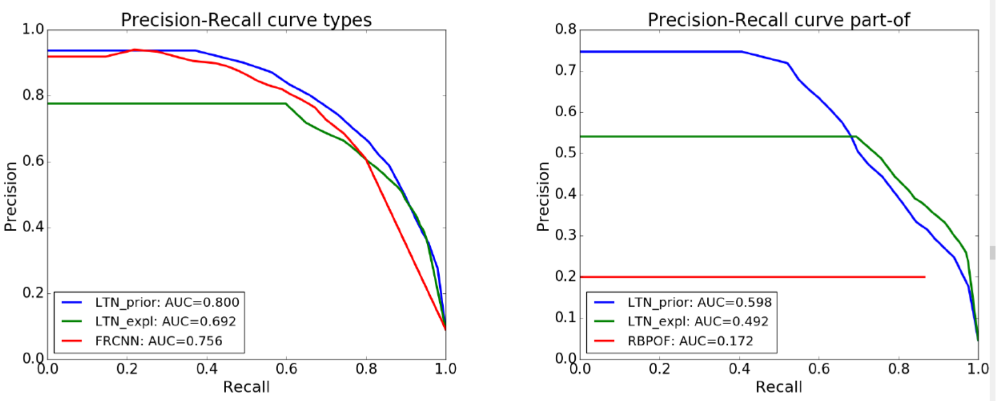


Figure 3.8: Precision-recall curves for indoor objects type classification and the *Part Of* relation between objects. Image of [2]

The results show that for the types of objects and for the "Part of" relation, LTN trained with prior knowledge provided by mere logical axioms perform better than LTN trained only with examples, see Figure 3.8. In addition, LTN was trained using the results of Fast R-CNN. This has meant that previous knowledge of LTN improves the performance of the Fast R-CNN (FRCNN) object detector.

This is because LTN makes the choice to consider the semantics and geometry of the data. This makes the LTN classifier robust with a drop in precision, but the logical axioms make up for this drop. Therefore, they concluded that the experiment indicates that the LTN axioms offer robustness when dealing with noise.

Regarding the fall, it is understandable to expect that there will be a negative impact on performance. But one can see a growing difference between the drop in performance of LTN trained with examples only and LTN trained including prior knowledge. Next, we show how to code a set of K formulas in LTN so that each formula in $\mathscr{K}$ has an LTN, and then, they have aggregated all networks with an operator and according to Equation (6).

Starting from an example where $\mathscr{K}$ is derived by the formula $drive(x,y) \to Vehicle(y)$, with input elements v and u. The grounding calculation is represented in the LTN in Figure 3.9, with $\mathscr{G}(x) = v, \mathscr{G}(y) = u$ and $k = 2$, therefore $\mathscr{G}(drive(v,u) \to Vehicle(u))$. The parameters to be learned are $\mathscr{W}_d, \mathscr{V}_d, b_d, u_d, \mathscr{W}_V, \mathscr{V}_V, b_V, u_V$ (where d means an inverter and V means the Vehicle).

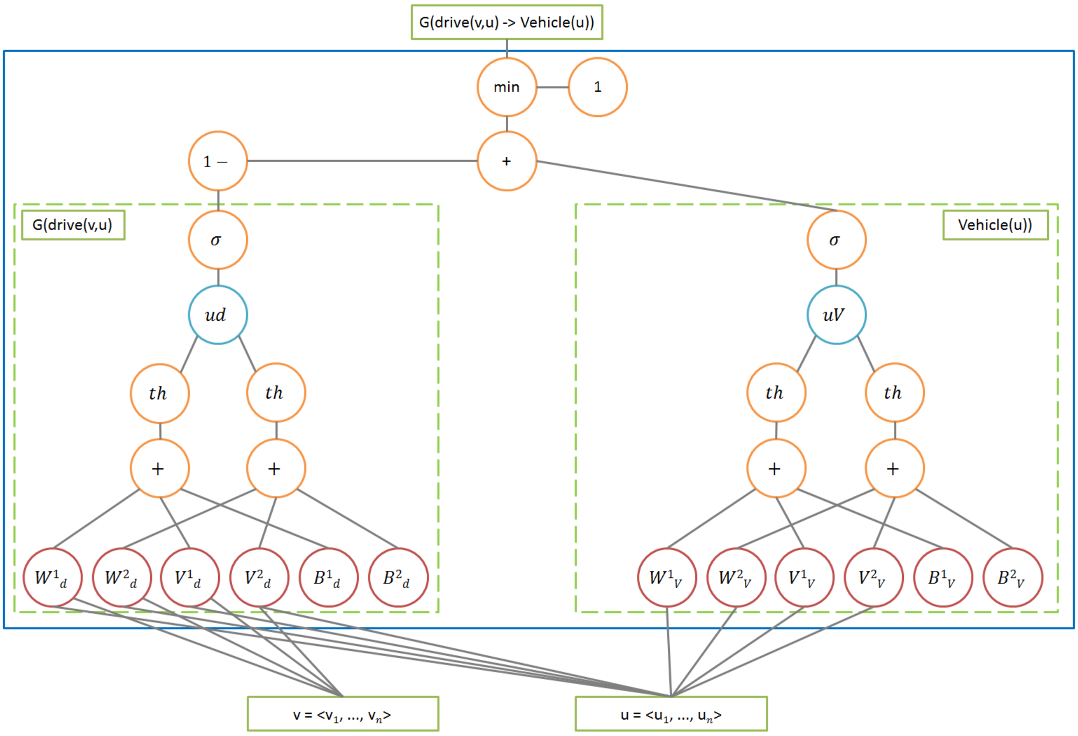


Figure 3.9: LTN for the formula $drive(x,y) \to Vehicle(y)$, $TensorFlow^{TM}$ to $drive(x,y)$ and other to $Vehicle(y)$. Image adapted from Donadello [2]

In this way, we understand that learning is accomplished by maximizing the true value, degree of satisfaction, of the formula $drive(x,y) \to Vehicle(y)$, see Definition 5. However, a knowledge-base $\mathscr{K}$ can contain many formulas: $\mathscr{K} = \phi_1, \ldots, \phi_q$ and the ground $\mathscr{G}(\phi)$ is calculated for each formula $(\phi \in \mathscr{G})$ that obtains a set of LTNs.

The blocks of the networks, represented by dashed rectangles in Figure 3.9, correspond to predicates in the $\mathscr{P}\mathscr{L}$ that can appear in various formulas. In other words, some blocks can be linked to blocks of other formulas and a very complex network will be formed.

In the last stage, the grounding $\mathscr{G}(\phi)$ (the outputs), with $\phi \in \mathscr{K}$, will be connected to some operators the implementation of $\wedge$ that will be defined according to the chosen t-norm. This will return the grounding $\mathscr{G}(\mathscr{K})$ of the entire knowledge base $\mathscr{K}$.

3.4 YOLO

YOLO, [3], is a system that provides the delimitation of boxes using dimension groups as anchorage boxes. The YOLO framework will substitute in our method the Fast-R CNN from Donatello's experiment. This is because YOLO currently has the best performance for classifying and identifying objects in images in a generic way, being ideal for the purpose of this research to have data from only one niche, in this context, soccer games.

This framework has the behavior defined with the cell and is identified by 4 coordinates for each bounding box, represented by: *tx, ty, tw, th*, for tx and ty the top left corner of the image by (cx,cy) and for width and height by tw and th. The predictions in YOLO correspond to, see 3.10:

$$
\begin{aligned}
b_x &= \phi(t_x) + c_x \\
b_y &= \phi(t_y) + c_y \\
b_w &= p_x e^{t_w} \\
b_h &= p_h e^{t_h}
\end{aligned}
$$

So it is possible to identify if the cell is offset from the upper left corner of the image by (c_x, c_y) and the previous bounding box has width and height p_x, p_h. During training, the sum of error loss squared, Redmon [3] explains how this is used, as the truth on the ground when determining a prediction from the coordinates is determined as $t*$. Then the gradient will be the value of the truth on the ground, obtained by subtracting the given prediction, so: $\hat{t}*$ - $t*$. This ground truth value can be easily calculated by inverting the above equations.

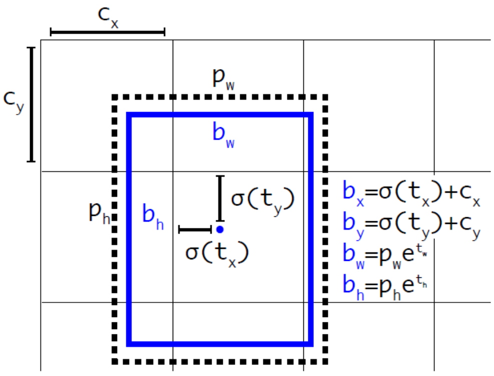


Figure 3.10: Bounding boxes with dimension priors and location prediction. Image from Redmon [3]

Each box predicts the classes the bounding box may contain using multi-label classification and independent logistic classifiers. During training, there is binary cross-entropy loss for the class predictions. According to Redmon [3] this occurs because the classes can be moved to more complex domains like the Open Images Dataset.

Therefore, YOLO training is done on full images without any difficult negative mining. According to Redmon [33], the neural network structure used in the framework is Darknet and its training took place at various scales, such as data augmentation and batch normalization.

3.5 Spatial-Temporal Reasoning

Studies on Spatial-Temporal Reasoning (STR) generally use visual data, and focus on the analysis of interactions with objects, observing characteristics of the environment through videos. According to Tayyub [34], in the analysis of spatial temporal relations, the qualitative approach is usually more successful, as it captures the main spatial and temporal changes in visual data and has become quite common in the representation of activities.

Such a study should not be based on just one dimension, as its characteristics may change over time and something in the *present* may be justified by a *previous state*. That is why it is important to obtain a representation that allows the analysis of the spatial and temporal dimensions in the same model. Since the states of an object are directly connected to its characteristics.

3.5.1 Qualitative and Quantitative Spatial-Temporal Relation

By using visual data to analyze interactions between objects, the qualitative and quantitative spatial-temporal relation (QQSTR) captures important changes in instances over time. The choices of these characteristics in the framework were defined by the great representativeness that this information describes in the analyzed visual data, explains Tayyub [34]:

Qualitative implies that two objects are partially overlapping, without specifying how much overlap exists, such as describing when an activity begins and ends before another activity begins.

Quantitative unlike the qualitative, is defined as something measurable in quantity, such as in describing that two objects, for example: overlap by 30%.

Spatial describes property and relations between objects that exist in space, such as poses of objects, poses of objects relative to other objects, the direction of absolute and relative movement, etc.

Temporal describes the properties or relations of objects or activities over time, like interval algebra. Identifying an activity's start time, its duration and completion, and whether there was overlap or relations with other properties, objects, or activities.

These features are used by the QQSTR method with each feature responsible for the complete characterization of an action F = <F1, F2, F3 >, as follows:

F1: Qualitative Spatial: A histogram is used to identify subsequences, that is if the sequence repeats it is suppressed until a change occurs. Suppressing the minimal blocks will suffice to describe an action.

The spatial relation between pairs of objects can be classified as follows, see Figure 3.11, with their representations being: discrete, partial or partial overlap, inverse part, or equality.

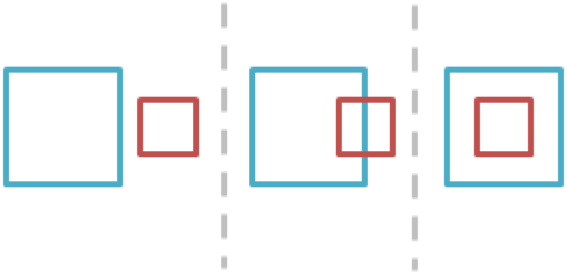


Figure 3.11: Each layout is categorized into F1, left - discrete D, in the center - PO partial overlap, and right - O overlap, inverse part, or equality. Image to illustrate the QQSTR, method of Cohn [4]

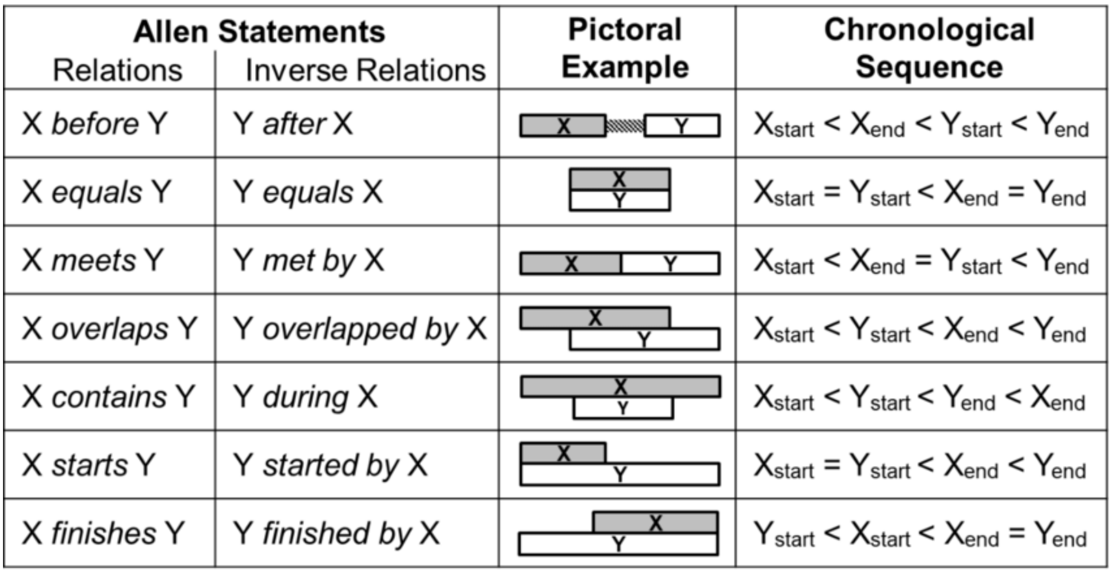
F2: Qualitative temporal In F1 it is not possible to have the notion of time in an event. Usually this description is made by Allen's interval algebra, cited in [23], but it does not encode quantitative relations of duration.

Allen's Interval Algebra is an approach to reasoning about time using the notion of time intervals and binary relations [35]. A *time interval* T is an (X, Y) where X is smaller than Y and both are points in time. The relation between these two points can be described by one of the relations in Table 3.1.

So in Tayyub [34] the relation of encounters is added as a qualitative measure according to a relative duration between two different and consecutive spatial relations. The temporal classifications of the duration of spatial relations can be short, equal, or long, as shown in Figure 3.12.

F3: Spatial quantitative: This representation is used to help the model recognize the different activities that resemble the qualitative representation. This uses the Euclidean

Table 3.1: Relations in Allen's Interval Algebra. Image from Mate [5]



Allen Statements		Pictoral Example	Chronological Sequence
Relations	Inverse Relations		
X *before* Y	Y *after* X		$X_{start} < X_{end} < Y_{start} < Y_{end}$
X *equals* Y	Y *equals* X		$X_{start} = Y_{start} < X_{end} = Y_{end}$
X *meets* Y	Y *met by* X		$X_{start} < X_{end} = Y_{start} < Y_{end}$
X *overlaps* Y	Y *overlapped by* X		$X_{start} < Y_{start} < X_{end} < Y_{end}$
X *contains* Y	Y *during* X		$X_{start} < Y_{start} < Y_{end} < X_{end}$
X *starts* Y	Y *started by* X		$X_{start} = Y_{start} < X_{end} < Y_{end}$
X *finishes* Y	Y *finished by* X		$Y_{start} < X_{start} < X_{end} = Y_{end}$

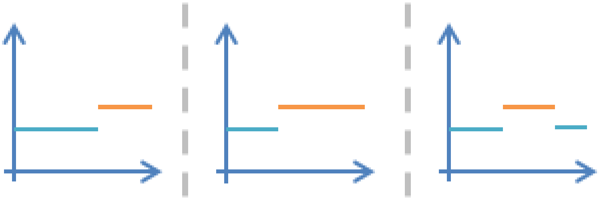


Figure 3.12: The scheme of each layout categorized into F2, left - short, in the center - equal, and right - long. Cohn's [4] adapted image

distance through the centroids of the identified objects, or the relative direction of motion. Contextualizing, thus, the previous characteristics in a measurable, quantitative way.

Qualitative Reasoning is not just a representation of the physical world, according to Cohn [4] it is also important in this abstraction as it is used for predictive models, diagnoses, and explanations of environments. By introducing physical characteristics, we have Qualitative Spatial Reasoning (QSR). This feature allows to obtain representation and reasoning with dimensional spatial entities, that is, kinematic features.

Thus the model has applications in areas such as geographic information systems, robotic navigation, common sense about physical dimensional situations, spatial preposition semantics in natural language, and, to which this research applies, visual language syntax and semantics, as explains Cohn [10].

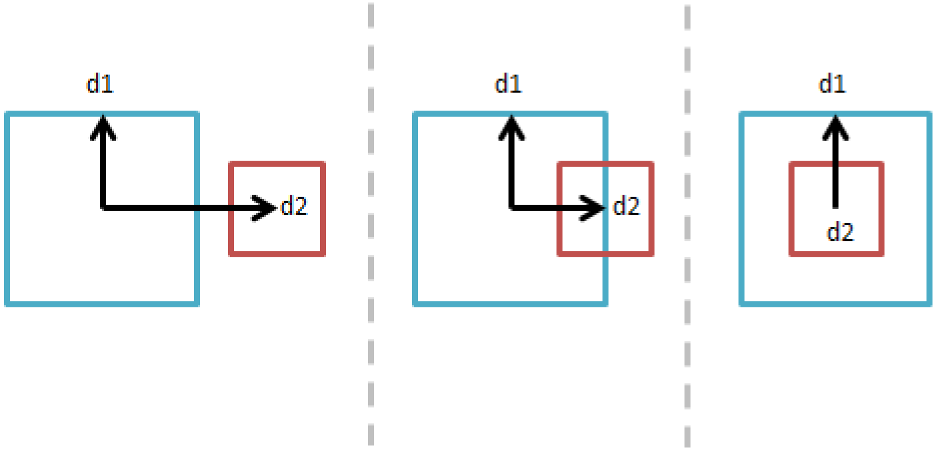


Figure 3.13: Demonstration of the difference between the centroids of objects in a quantitative way. Cohn's [4] adapted image

3.6 Semantic Video Interpretation

During the course of the research, we must pay attention to the object under study and its main characteristics for its analysis [24]. LTN can analyze information about spatial features as it can extract it from the provided visual data. However, in a video, LTN would analyze each frame of the video individually, without analyzing the context of the temporal information.

But when we analyze videos, we have a new dimension of characteristics, *time*. For this reason, spatial-temporal reasoning is necessary for this approach. That's why we propose to adapt the LTN so that each frame is analyzed individually to semantically describe the set of frames by which the video is formed.

Recognition of video semantics is a fundamental research problem in computer vision and multimedia analysis. Video data is usually represented by high-dimensional features. Zhang [36] identifies that semantic video recognition performance may deteriorate, due to irrelevant and redundant components included in high-dimensional representations.

Thus, it is important to emphasize a deeper understanding of video activities. This goes beyond the recognition of underlying concepts such as actions and objects, but also the construction of semantic representations, being profound it requires reasoning about the semantic relations between these concepts.

The interpretation of video content consists of the construction of a semantically coherent composition of basic (atomic) elements of knowledge. According to Aakur [37] they are called *concepts detected in the videos* which represent the individual actions and objects needed to form an interpretation of an activity.

3.6.1 Identification of objects and their relations in images and videos

We call an *object* any and all elements identified in a scene, be it a person, a member of a body, a kitchen utensil or animal. To carry out the identification of an object in a video frame, or in an image, we need to clearly identify the objective of the model we are dealing with. Because information about object placement and description may be sufficient for a search, while colors, dimensions, and relations with other objects may be relevant for another objective, as Cohen[24] mentions.

Through the identification of elements in an image, it is possible to describe and infer situations that an image represents, already in a video, the understanding of information is related to the events that are presented, events between objects, how they occur, and their consequences. And to analyze each context and purpose of descriptions, some characteristics must be analyzed.

3.6.2 Activity Graph

Classification of object actions in videos can be learned by systematizing the information in Activity Graphs, as shown by Srichar [38]. They are graphs used specifically to model objects and the relations between them over time.

When more than one event occurs in the same frame set in a video, it is called a *complex scene*. This scene can be analyzed using *tracks*, which will last as long as there are spatial relations between certain pairs of objects as illustrated, for example, in Figure 3.14:

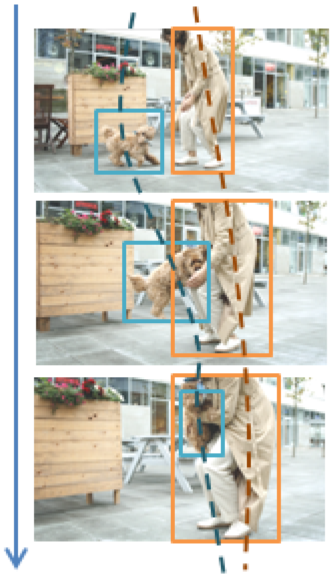


Figure 3.14: Example of tracks between two object relations

Each change in the spatial relations between a pair of objects is called an *episode*. As example, one can describe Figure 3.14 as follows: In the first episode, the dog $o1$ and the woman $o2$ have the relation $r1$ in the time interval $I1$. Their first representation in this model will be like (1) $holds(dog, woman, r1, I1)$; in the second episode, the identified objects have a new relation $r2$ in $I2$ described like this (2) $holds(dog, woman, r2, I2)$; finally, in the third episode, we have the last relation $r3$ in $I3$, $holds(dog, woman, r3, I3)$. After describing the spatial relations, we can insert the relations between the time intervals that were indicated, as follows: *meets(I1, I2)*, *meets (I2, I3)* and *before (I1, I3)*.

As concerns the episodes, it is necessary to describe each episode for all pairs of objects in the video, to build the level 0 of the activity graph. Sridhar [38] mentions, that the graph will be constructed with vertices, like an episode, and the edges of Allen's time intervals are described with temporal relations: *before, meets, overlaps, starts, during, ends, and equals*.

Since the episode is described as: "in frame *f* there is a spatial relation *r* between objects *o1* and *o2*", in the logical form is: $holds(r(o1, o2), f)$ and formalized by the quadruple: E = $<o1, o2, T, r>$, containing the objects $o1$ and $o2$. The consecutive sequence of frames (f1 + f2 + ... + fn) was a repeated relation that occurs in these frames.

After the elaboration of the complete graph, an *attention* mechanism is used to change the level of abstraction of the graph to level 1. This allows the identification of "super events" later allowing the classification of activities in a generalized way [38].

This happens by grouping episodes for vertices and the temporal relations between these groupings will be the edges. When different actions occur in the same frame, the relation between them is *during*, being called *objects in the foreground*, those that start the analysis of the frame. The other objects comprise the *objects as the background*.

Chapter 4

Proposed Model

Despite the advantages highlighted in the previous chapter, LTN identifies the *PartOf* relation between pairs of bounding boxes[2] in static concepts. For this reason, there is such an accurate performance and applied only to analyzing images. That is, the advantages it presents are superior to other models that also identify objects in images. But this evidence is a hypothesis for this research as we explore the application in a semantic analysis of videos.

Given this context, to use LTN in video analysis, its syntactic and semantic descriptive model must be extended. This way, LTN is able to have its advantages applied in spatial-temporal relations of video descriptions. This approach forms the basis of this work to evaluate the advantages of neural symbolic reasoning in the descriptive analysis of videos.

Our adaptation proposal is to use the YOLO framework to process the video in a few steps so that the problem is treated according to the "divide and conquer" approach. Thus, it would be possible to reduce the video complexity for individual frame analysis for LTN applications. Such adaptation would take place as follows:

- The input is a video that is initially divided into frames, i.e., sequences of images
- Chooses a frame at each selected number interval
- Identifies the objects in the selected image and using dimension clusters as anchor boxes, the network predicts 4 coordinates for each bounding box
- Classifies each object by predicting an objectivity for each box boundary using logistic regression

This output is handled according to the consistent representation of time and space provided by QQSTR, described in Tayyub's [34] approach. Such a representation provides a way of organizing information about time and space by its qualitative and quantitative characteristics.

Finally, the Activity Graph model of Sridhar [38], is still needed. Because from it, it will be possible to reason the spatial relations with the temporal relations identified throughout the process.

When we approach this last step in more detail, we will have a clearer view of the importance of using symbolic neural reasoning, through LTN, to identify the spatial relations of the data provided.

Thus, we emphasize the importance of these steps. Because through them it is possible that we are not restricted to the information retained in the videos, as we use logical and symbolic data that can, in turn, be endowed with inferences, reasoning, and logical rules.

In addition, any type of information that may be related to evolution over time can be analyzed with our proposed approach. As mentioned in the first chapter, data that evolve over time can be monitoring medical records, analyzing social behavior, identifying climate changes, and analyzing the stock market in a period, among others.

4.1 The Extension Proposed

We can use the neural symbolic approach to describe videos. Through a few steps, it is possible to adapt the information about videos to use the LTN framework to identify spatial relations. Relations of object pairs are identified and classified in video frames by the YOLO framework. For this, the actions that occur during the video are represented based on the approach of the activity graph. To explain this proposed approach, we represent the steps illustrated in Figure 4.6:

1. YOLO has as input a video and output base information with the coordinates of the pairs of bounding boxes and the classification of their respective objects;

2. LTN receives output from YOLO and identifies spatial relations between pairs of bounding boxes;

3. Construction of the Activity Graph where the edges are the temporal relations, based on Allen's algebra, and the nodes are the grouping of frames with the same spatial relation in related time;

4. Perform the reasoning for the video description with the modeled information.

We start from the principle expounded by Badreddine [27] which shows that the standard approach to the multiple-label problem is to provide explicit negative examples for each class.

However, LTN can use its prior knowledge to relate classes, making it a powerful tool in the case of the multiple-label problem, when labeled data is usually scarce.

And these superior characteristics of LTN are realized through the factors that make up its structure, which are presented below.

Domain *labels* denoting the relations

Constant are the *bounding boxes*

Predicates spatial relations of bounding boxes pairs

Axioms represent the mutual exclusion of labels on spatial relations. As a result, negative examples are not used explicitly in this specification

Grounding $\mathscr{G}$(itens) = $\mathscr{N}^{\triangle}$ vectors are used to represent class labels

Learning As before, the fuzzy logic operators and connectives are approximated using the stable product configuration

Querying LTN constraint learning

By joining the spatial relations described in LTN with the qualitative spatial characteristics of QQSTR, we had spatial relations that cover several cases. They arise from 3 labels capable of distinguishing 24 spatial relations, shown in Figure 4.1. New spatial relations, based on QQSTR on the background knowledge of the LTN, allow identifying different positions in relation to the pairs of objects identified in each image.

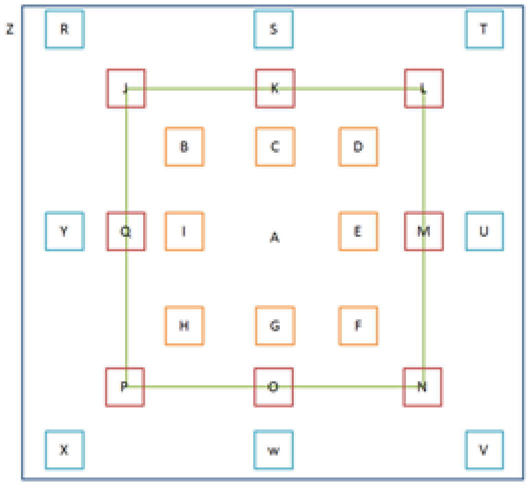


Figure 4.1: Types of spatial relations based on QQSTR and LTN

$inSideLeftAltAbove(B,A) \leftarrow p(A,B), left(B,A), above(B,A)$

$inAltAbove(C,A) \leftarrow p(A,C), above(C,A)$

$inSideRightAltAbove(D,A) \leftarrow p(A,D), left(A,D), above(D,A)$

$inSideRight(E,A) \leftarrow p(A,E), left(A,D)$

$inSideRightAltBelow(F,A) \leftarrow p(A,F), left(A,D), above(A,F)$

$inAltBelow(G,A) \leftarrow p(A,G), above(A,G)$

$inSideLeftAltBelow(H,A) \leftarrow p(A,H), left(H,A), above(A,H)$

$inSideLeft(I,A) \leftarrow p(A,I), left(I,A)$

$edgeSideLeftAltAbove(J,A) \leftarrow po(A,J), left(J,A), above(J,A)$

$edgeAltAbove(K,A) \leftarrow po(A,K), above(K,A)$

$edgeSideRightAltAbove(L,A) \leftarrow po(A,L), left(A,L), above(L,A)$

$edgeSideRight(M,A) \leftarrow po(A,M), left(A,M)$

$edgeSideRightAltBelow(N,A) \leftarrow po(A,N), left(A,N), above(A,N)$

$edgeAltBelow(O,A) \leftarrow po(A,O), above(A,O)$

$edgeSideLeftAltBelow(P,A) \leftarrow po(A,P), left(P,A), above(A,P)$

$edgeSideLeft(Q,A) \leftarrow po(A,Q), left(Q,A)$

$outSideLeftAltAbove(R,A) \leftarrow d(A,R), left(A,R), above(R,A)$

$outAltAbove(S,A) \leftarrow d(A,S), above(S,A)$

$outSideRightAltAbove(T,A) \leftarrow d(A,T), left(A,T), above(T,A)$

$outSideRight(U,A) \leftarrow d(A,U), left(A,U)$

$outSideRightAltBelow(V,A) \leftarrow d(A,W), left(A,V), above(A,V)$

$outAltBelow(W,A) \leftarrow d(A,W), above(A,W)$

$outSideLeftAltBelow(X,A) \leftarrow d(A,X), left(X,A), above(A,X)$

$outSideLeft(Y,A) \leftarrow d(A,Y), left(Y,A)$

$in(A,Z) \leftarrow p(Z,A)$

The spatial relations presented above can be divided into three categories of labels: O - overlapping, PO - partially overlapping, and D - discrete. This definition is given according to the coordinates identified by YOLO in the delimitation of the identified objects. In the next chapter, we describe in more details how these spatial relations influence the reasoning of the video description.

4.2 Syntax and Semantics

This section describes the syntactic and semantics addressed in our hypothesis. It is carried out according to LTN formalities, facilitating the integration of all the methods used.

The knowledge calculated by the LTN is called *grounding* ($\mathscr{G}$). In our case, grounding refers to the spatial relations of the selected frames. And these relations are always referring to bounding box pairs like this: *relation type* (b,b'), where b and b' are bounding boxes identified in an image and the *relations* between them.

We consider that, as in LTN, the foundation of $\mathscr{G}$ i in a first-order language PL, so it is necessary to obey the functions of the PL signature that satisfies the conditions:

Definition 1 *Also in LTN, grounding $\mathscr{G}$ for a first-order language PL is a function of the PL signature that satisfies the conditions:*

1. $\mathscr{G}(c) \in \mathbb{R}^n$ *for every constant symbol* $c \in C$ *that is bounding box* $b \in Pics$;
2. $\mathscr{G}(f) \in \mathbb{R}^{n.a(f)} \to \mathbb{R}^n$ *for every functional symbol* $fin\mathscr{F}$;
3. $\mathscr{G}(P) \in \mathbb{R}^{n.a(P)} \to [0,1]$ *for every predicate symbol* $P \in \mathscr{P}$ *that are* $P_1 and P_2$

Grounding can be done by methods using predicates or by logical rules. We start by introducing grounding by logical rules. The classification of a bounding box occurs through a score function σ of all possible classes in P_1 on a single bounding box, and its grounding also has its coordinates from upper left and lower right points:

$$< score(C_1,b),...,score(C_{|P1|},b),x_0(b),y_0(b),x_1(b),y_1(b) > \mathscr{G}(C(b)) = \sigma(score(C,b))$$

The groundings for the spatial relations are membership functions, which return the degree of membership of an element. Fuzzy Logic represents the true value of an atomic formula.

The following relations are defined from the Cartesian position of the vertices that compose the bounding box of the objects identified by x,y,w, and h. Being x and y the ordered pair of the lower-left vertices of each box, w the width, and h the height.

Let $b,b' \in C$ be two bounding box constants, and β be the angle made by an angle in a unitary circle between the center of the circumference, that will be the centroid of b, and a point on the circumference, that will be the centroid of b'.

Below is a table that summarizes the table in Appendix A. In this table, we have the semantic definitions for the new spatial relations proposed in this manuscript in Table 4.1. Settings include the placement of related bounding box points and the angle between their centroids.

Table 4.1: Summary information from the table in the appendix

Group	Spatial relation Example	Definitions
b with limitations within of $b^{'}$	$\mathscr{G}(inSideLeftAltAbove(b,b^{'}))$	$x_0(b) \geqslant x_0(b^{'})$ $y_0(b) \leq y_0(b^{'})$ $x_1(b) \leq x_1(b^{'})$ $y_1(b) \geqslant y_1(b^{'})$ $regardlessofangle$
b with limitations that overlap the edges of $b^{'}$	$\mathscr{G}(edgeAltAboveSideLeft(b,b^{'}))$	$x_0(b) < x_0(b^{'})$ $x_1(b) < x_1(b^{'})$ $y_0(b) > y_0(b^{'})$ $y_1(b) < y_1(b^{'})$ 135^o
b with limitations outside the $b^{'}$	$\mathscr{G}(outSideRight(b,b^{'}))$	$x_0(b) < x_0(b^{'})$ $x_1(b) < x_1(b^{'})$

4.2.1 Proposed Spatial Relations

For the evolution of all relations mentioned above we have as a basis the following relations used in our approach.

After some adaptations, it was defined that the classifications would be according to the coordinates to be defined as $h1$ and $w1$ height and length, $(x1,y1)$ as the ordered pair of the lower-left vertical of bounding box 1 and $h2$ and $w2$ height and length, $(x2,y2)$ as the ordered pair of the lower-left vertical of bounding box 2:

considering as "O":
$if(((x2 >= x1)and((x2+w2) <= (x1+w1)))or((y2 > y1)and((y2+h2) < (y1+h1))))$

considering as "PO":
$if(((x2 < x1)and((x2+w2) > x1)and$
I . case 1
$(((y2 < y1)and((y2+h2) > y1))or((y2+h2) > (y1+h1))))$
II . case 2
$or(((x2 < x1)and((x2+w2) > x1)and((x2+w2) < (x1+w1)))and$
$((((y2+h2) > (y1+h1))and(y2 < (y1+h1)))or((y2 < y1)and((y2+h2) > y1))))$
III . case 3
$or(((x2 > x1)and((x2+h2) < (x1+h1)))and$
$((((y2+h2) > (y1+h1))and(y2 < (y1+h1)))or((y2 < y1)and((y2+h2) > y1))))$
IV . case 4
$or(((x2 > x1)and(x2 < (x1+w1))and((x2+w2) > (x1+w1)))and$

$((((y2+h2) < (y1+h1))and((y2+h2) > y1))or((y2 > y1)and(y2 < (y1+h1)))))$

considering as "D":
$if(((x2+w2) < x1)or(x2 > (x1+w1))or(y2 > (y1+h1))or((y2+h2) < y1))$

From these definitions, we analyze LTN which identifies the input bounding boxes have. Within a toy example, the classification follows as a general context as described in 4.2.2.

However, in chapter 5 we address the classification of the action for the specific context of the experiment, soccer. This happens based on the same generic descriptions and thus demonstrating the efficiency of the method with all types of temporal data, that is, that evolve over time as discussed above.

To demonstrate LTN constraint learning overtime during learning, let's use an example. In this first moment we manually analyze the video *dogandwoman.mp*3 (attached) to defend the hypothesis presented, outlined in Figure 3.5.

The first step is to submit the video to the YOLO framework, to get the selected frames that will be passed on to the LTN framework. However, before submission, it is necessary to build the background knowledge in the LTN.

The following is a diagram of the first step being YOLO identifying and classifying the bounding boxes of the images:

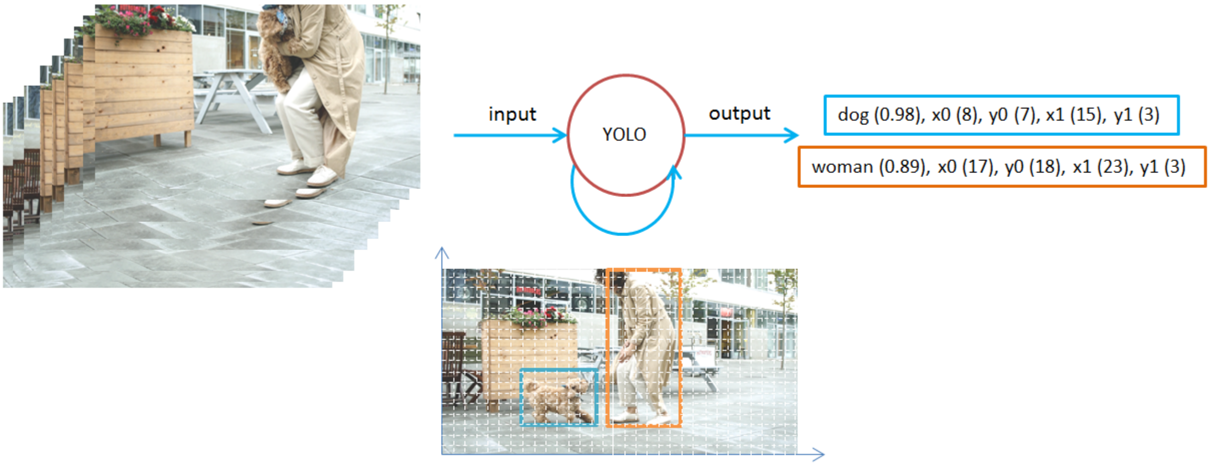

input
YOLO
output
dog (0.98), x0 (8), y0 (7), x1 (15), y1 (3)
woman (0.89), x0 (17), y0 (18), x1 (23), y1 (3)

Figure 4.2: Example of notation made from YOLO

The categorization goes through stages, analyzing the score of the ratings to indicate the object of that bounding box 4.2:

BB1: <$dog(0.98), horse(0.02), woman(0), child(0), x0(8), y0(7), x1(15), y1(3) >$
BB2: <$dog(0), horse(0), woman(0.89), child(0.16), x0(17), y0(18), x1(23), y1(3) >$

According to the positions of the bounding box pairs, we can analyze which of the spatial relations as illustrated in Figure 4.1 we can identify between them. Then we group the frames that have the same description of the spatial relation between the identified objects. In this example, we have 3 groups shown in Figure 4.3:

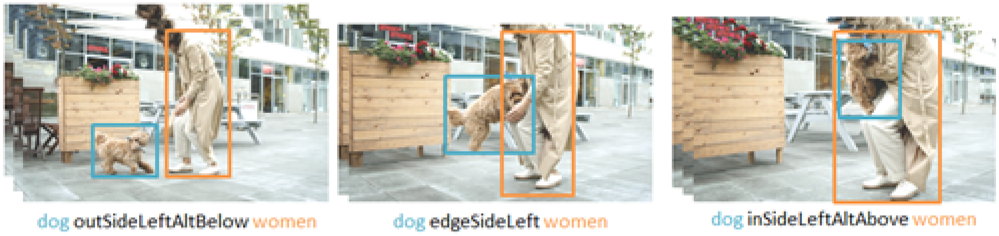


Figure 4.3: Separation of frame groups from objects with the same spatial relation

Identification of the beginning and end of the spatial relations that the objects maintain among themselves during the video 4.4.

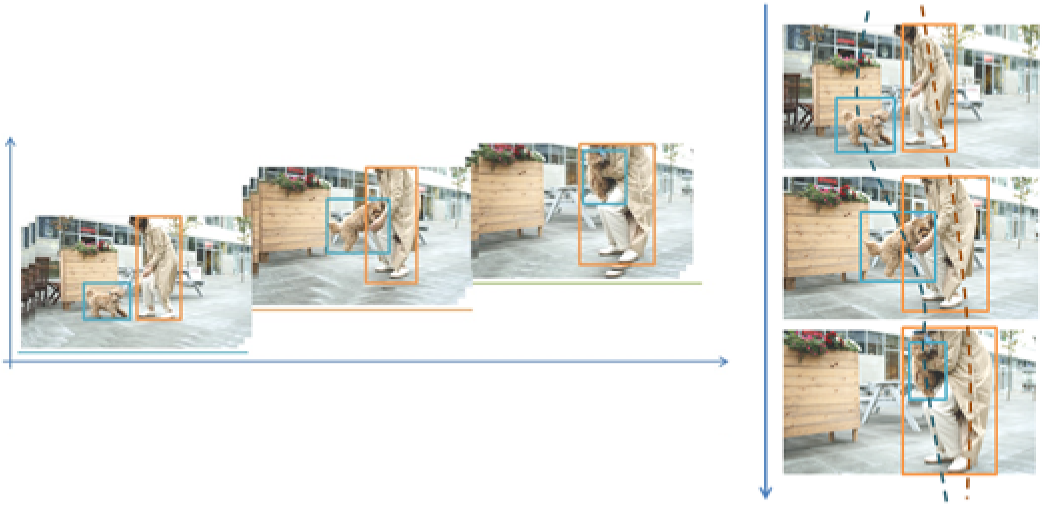


Figure 4.4: Identification of object tracks during the video

The graph construction has each episode, as a grouping, with the edges as its temporal relations 4.5.

Through the activity graph, we have the spatial-temporal relations extracted from a video with a logical description. Thus being able to perform logical reasoning about it, in addition to the semantic description mentioned in Table 4.2:

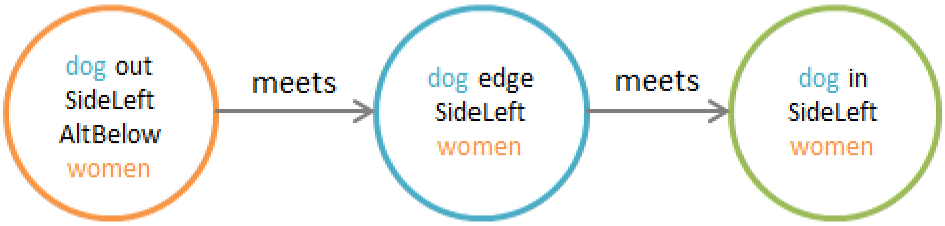


Figure 4.5: Construction of the activity graph

Episode 1	holds(dog,woman,outSideLeftAltBelow,I1)
Episode 2	holds(dog,woman,edgeSideLeft,I2)
Episode 3	holds(dog,woman,inSideLeftAltAbove,I3)
	meets(I1,I2) and meets(I2,I3) and before(I1,I3)

Table 4.2: Spatial-temporal relations of toy example

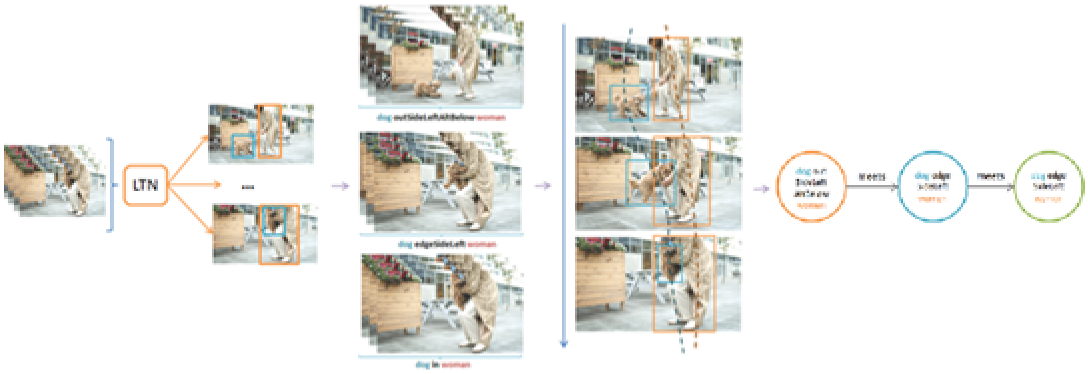


Figure 4.6: Approach hypothesis scheme

4.2.2 Inferences and Learning About Actions of the Toy Example

The inferences about a description of the episodes must observe the changes in spatial relations and the temporal relations between them. In terms of a neural-symbolic approach, we can use background knowledge combined with logical constraints, for learning actions. Let's see some cases based on the toy example presented:

- *pick_up*: from "out" to "in"
  (holds (o1, o2, out, I1) and holds (o1, o2, in, I2) and (before (I1, I2) or meets (I1, I2)))

- *leave*: from "in" to "out"
  (holds (o1, o2, in, I1) and holds (o1, o2, out, I2) and (before (I1, I2) or meets (I1, I2)))

- *went_up*: from “below” to “above”
  (holds (o1, o2, below, I1) and holds (o1, o2, above, I2) and (before (I1, I2) or meets (I1, I2)))

- *go_down*: from “above” to “below”
  (holds (o1, o2, above, I1) and holds (o1, o2, below, I2) and (before (I1, I2) or meets (I1, I2)))

- *got_up*: *pick_up* and *went_up*

- *let_go*: *leave* and *go_down*

In the example presented, we find in the first episode that the dog is “out", "left" and “below" about the woman, then in the last episode it is “in", continues “left", that is, it fits the case *got_up*.

The great advantage of this approach is that one can detect complex activities based on simpler activities. Approaches that use end-to-end Machine Learning models need an absurdly large number of examples, while what we propose resembles something more like the human learning process, with few examples.

By abstracting the visual information from a video that demonstrates the heartbeat of a patient undergoing a surgical procedure, let us imagine a situation where a sequence of heartbeat frequencies are mapped to what kind of action is required.

We could classify if it is in a normal, slow or fast state, and identify the action that this sequence of states could represent. And in a future work, predict which actions could be taken according to the current state by analyzing the history already covered during the video.

As described so far, this approach considers that the data evolving in space and time characteristics can be reasoned through a neuro-symbolic approach in the classification of their spatial relations represented in an activity graph.

4.3 Representation of information

This section presents a generalized formalization of each step taken to extract the information from the video till its logical representation. To achieve this, we present the logical formalization used 4.1, followed by the algorithm of information extraction, and finally an example of how this execution occurs.

The formalization for logical representation used in this work is based on [2], which consists of declaring unary and binary predicates, the unary ones representing the category of the object identified in the video and the binaries representing the spatial and binary predicates' temporal relations. The identification and categorization of objects is done using the YOLO framework, while the identification of spatial relations is done using LTNQSTR 4.1.

4.3.1 Description of Background Knowledge

Based on LTN formalization, we also have a knowledge base defined by <P,C>, being predicted of two types, unary P1 (object classification) and binary P2 (spatial and temporal relations). Some examples for the predicates would be:

P1 - Person, Dog, Car, Ball, Notebook... P2 - [24 spatial relations, described in 4.1], [Allen's temporal algebra, in 3.1]

The object classification is obtained as knowledge prior to learning, as well as the spatial relations that were defined in the 4.1 section and the temporal relations described by Allen's algebra, in Y.

4.4 Algorithm

This section presents the formalization of each step to extract the information from the video till the logical representation. For this, we present the logical formalization used, followed by the information extraction algorithm, and finally an example of how this execution occurs.

The formalization for logical representation used is based on Donadello [2], mainly in its division of unary and binary predicates.

The identification and categorization of the objects is done done through the YOLO framework (Section 3.4), while the classification of the spatial relations is done through LTNQQSRT (Section 3.5.1); the identification of the temporal relations is based on Allen's Algebra [23].

The knowledge base is defined by <P, C >predicates and constants, the predicates being of two types, P1 unary and P2 binary. For example, for unary predicates we have: *Person, Dog, Car, Ball, Interval*, and for binary predicates: *out_left_below, in_left_below, before, after*.

We also have predicates that help in the space-time representation, including *pair, interval* and *holds*, which represent respectively: the pairs of objects of the formalized instance, declaration of the interval, and finally, the relation between the predicate of the spatial relation and the instance of interval.

As described in Section 3.4, object classification is obtained as knowledge before learning, as well as the spatial relations that were defined in section 3.6 and the temporal relations described by Allen's algebra, in 3.5.1. Our logical representation needs to identify (1) the objects that are being analyzed, (2) each spatial relation that occurs in the video, and (3) the temporal relation between these instances. For this we propose the following representation:

$$\{Person(p1), Dog(d1), pair(p1,d1), Interval(i1), holds(out_left\_below(d1,p1),i1),$$
$$Interval(i2), holds(in\_left\_below(d1,p1),i2), before(i1,i2)\}$$

In this representation we have defined the objects present in the video:

$$Person(p1), Dog(d1), pair(p1,d1)$$

The predicates *Person/Dog* are the objects identified in the video, which are unary predicates. We use the binary predicate *pair* to identify which pairs of objects are being highlighted in this logical expression.

We then represent the spatial-temporal relation as:

$$Interval(i1), holds(out\_left\_below(d1,p1),i1)...before(i1,i2)$$

The attribute *interval* defines the time interval of the current instance, *holds* is the binary attribute for identifying the space-time relation, composed of the respective arguments: (1) spatial relation - this argument is a binary attribute consisting of the pair of objects in its relation, (2) temporal relation - is the identification of the time interval of this spatial relation. Finally, the binary predicate *before* is the temporal relation between the intervals of its arguments.

4.4.1 Algorithm Description

The algorithm can be divided into four main parts (1) Selection of Frames and Classification of Objects (2) Classification of Spatial Relations (3) Classification of Temporal Relations (4) Inference of the Action in the Video:

Preparing the video The video preparation consists of extracting selected frames from the video with the classification of its objects. For this (1) each frame of the video is saved in a folder, that is, the video is divided into several images. From this point (2) we select only a few frames for analysis, selecting images in the interval of 2 in 2 seconds. As it is not a stream transmission, but an analysis of a short video, we can drop some frames without compromising the result sought. After the sequential frame selection is done, (3) we submit these images to the YOLO framework to recognize and classify the objects present in the frames. The recognition by YOLO is highlighted with rectangles (or bounding boxes) the objects identified in each image, informing their position in the image, and the classification associated with it. At the end of this step, we can have a set of frames with their objects identified by their position in the image and classification.

Identification of spatial relations The identification of spatial relations occurs through the LTNQQSR framework, an example of implementation as a proof of concept, which receives as input an image with the identification of a pair of bounding boxes to identify the spatial relation between these two objects. These relations are for pairs of objects, identified through the position of the bounding box (bb) centroids and the relations between their boundaries, see 4.1, for example, $out_{l}eft_{b}elow(x,y)$. For this, we need a script that runs LTNQQSR for each frame processed by YOLO in the first step and this process results in the identification of spatial relations between two objects for each input image.

Construction of the activity graph and the logical expression The data structure chosen for such representation of the temporal space relations was an activity graph. In such a graph, each node is a spatial relation and the edges are the temporal relations section of the activity graph. For its construction, we (1) group the frames with the same sequential spatial relation, (2) represent each group as a node of the graph, (3) identify the temporal relations through Allen's algebra, and (4) join the graph nodes according to the identified temporal relations.

1. grouping consists of comparing the spatial relation frame by frame identified in step 2. As long as the spatial relation is the same, the sequential frames will be added to the same group. As soon as the analyzed frame presents a new spatial relation, we can create a new group for it and so on to return to the beginning of this process until the end of the frames;
2. each group is a graph node, for example, group 1 (frame 1 to 4) is the first node, group 2 (frame 5 to 8) is the second node;
3. the identification of temporal relations in our current scenario will always be *before / after*, as we deal with sequential actions and analyze object pairs per frame. However, this structure would support more complex relations with multiple object pairs and simultaneous actions;
4. we identify that we need to connect each node with its respective temporal relation.

Our logical expression is formed by three steps: (1) by the objects identified in the spatial relations in the node. These are unary predicates, and the identification of this pair of objects are represented by the predicate *par* (2) we have the declarations of the relations' spatial relations, where for each new group we have a new *Interval* followed by a *holds* that declares the spatial relation of the highlighted group, followed by the interval that corresponds to it. These temporal relations are made to represent each node of the graph or group of frames, and finally (3) we have the logical representation of the temporal relations, which are predicted as defined by Allen's algebra that identifies the relation between the intervals.

The following is the description of our algorithm 1:

Keep in mind that BB's can have 24 different spatial relations already presented in this chapter. Then we can infer the existing action by checking the actions identified by the nodes. In the next chapter, we present a practical example of formulating the rules and applying the hypothesis in the context of soccer games. As well as the inferences that occur to identify actions through logical axioms.

Algorithm 1 The algorithm name

```
procedure SELECTION OF FRAMES AND CLASSIFICATION OF OBJECTS($video$)
    $YOLOimages[\varnothing]$
    $frame \leftarrow first\_frame(video)$
    while $frame \neq NULL$ do                                      ▷ apply to all video frames
        $frame \leftarrow YOLO(frame)$                              ▷ apply YOLO for this frame
        $YOLOimages \leftarrow insert(frame)$                       ▷ save processed image
        $frame \leftarrow next\_frame(video)$                       ▷ next frame
    $returnYOLOimages$
procedure CLASSIFICATION OF SPATIAL RELATIONS($YOLOimages[]$)
    $imageSpatialRelation[\varnothing]$
    $image \leftarrow first\_image(YOLOimages)$
    while $image \neq NULL$ do                                      ▷ apply to all images processed by YOLO
        $spacial\_relation \leftarrow LTNQSR(image)$                ▷ apply YOLO for this frame
        $imageSpatialRelation \leftarrow (image, spacial\_relation)$  ▷ identified spatial relation with your image
        $image \leftarrow next\_image(YOLOimages)$                  ▷ next image
procedure CLASSIFICATION OF TEMPORAL RELATIONS($imageSpatialRelation[]$)
    $temporalClassification[\varnothing]$
    $image \leftarrow first\_image(imageSpatialRelation)$
    $pair \leftarrow next\_image(image)$                            ▷ next image of sequence
    while $image \neq NULL$ do                                      ▷ apply to all images processed by LTNQR
        $temporal\_relation \leftarrow Allen\_Interval(image, pair)$  ▷ identify relation temporal
        $temporalClassification \leftarrow (image, temporal\_relation, pair)$  ▷ associate the identified spatial relation with your image
        $image \leftarrow pair$                                     ▷ next image
        $pair \leftarrow next\_image(image)$                        ▷ next pair
procedure INFERENCE OF THE ACTION IN THE VIDEO($YOLOimages[], imageSpatialRelation[], temporalClassification[]$)
    $data = YOLOimages, imageSpatialRelation, temporalClassification$
    $apply\_inference(data, bk)$
```

Chapter 5

Experiment

One of the current challenges in AI is combining explicit symbolic knowledge in the form of rules with implicit sub-symbolic knowledge, such as the weights of a neural network. For this reason, we use LTN for such a challenge. It is a framework for a symbolic neural network using fuzzy logic for its reasoning of symbolic rules.

Thus, we propose a comparison between the behaviors of neural networks with and without symbolic rules in order to validate whether the combination of explicit and symbolic knowledge in the form of rules with implicit and sub-symbolic knowledge can excel in advantages over purely neural learning.

Following the scope of this investigation, our proposal for video interpretation constitutes of approaching steps for fragmenting the video into frames, identifying and categorizing the objects present in the video; identifying the spatial relation between pairs of objects to be treated, and logical inference about what action was performed in the video.

As we have several steps, we use the YOLO framework for the identification and categorization of objects on the frames; for semantics and definition of spatial relations, we adapted the QQSTR method within the LTN framework, and finally, for structuring the Spatio-temporal information of data, we used the Activity Graph.

For this approach we treat spatial relations as: *O totally overlapped, PO partially overlapped and D discretely non-overlapping*. For the present context and for the specific knowledge proposed, these summarized relations of those represented in the Appendix are sufficient for our purpose.

Demonstrating the importance of the abstract context of this approach, we present in this chapter each of the proposed algorithms and finally an application in a specific scenario of categorization and recognition of action types during a *soccer match*.

These spatial relations are built logically by checking the veracity of the following arguments, admitting BB1 and BB2: $BB1[x1,y1,w1,h1]$ and $BB2[x2,y2,w2,h2]$ with representatives: x - on the coordinate axis for the bottom left point of bb, y - on the abscissa axis from the bottom left point of bb, w - the width of bb, h - the height of bb:

Category	Case	Rule
O	1:	$[x2 \geq x1]\ AND\ [(x2+w2) \leq (x1+w1)]$
	2:	$[y2 \geq y1]\ AND\ [(y2+h2) \leq (y1+h1)]$
PO	1:	$[x2 \leq x1]\ AND\{(x2+w2) \geq x1\}\ AND\ [\{y2 \leq y1\}\ AND\ (y2+h2) \geq y1\ OR\ \{(y2+h2) \leq (y1+h1)\}]$
	2:	$[x2 \leq x1]\ AND\ \{(x2+w2) \geq x1\}\ AND\ \{(x2+w2) \leq (x1+w1)\}\ AND\ [\{(y2+h2) \geq (y1+h1)\}AND\ \{y2 \leq (y1+h1)\}\ OR\ \{(y2 \leq y1)\ AND\ (y2+h2 \geq y1)\}]$
	3:	$[x2 \geq x1]\ AND\ \{(x2+w2) \leq (x1+w1)\}\ AND\ [\{(y2+h2) \geq (y1+h1)\}AND\ \{y2 \leq (y1+h1)\}\ OR\ \{y2 \leq y1\ AND\ (y2+h2) \geq y1\}]$
	4:	$[x2 \geq x1]\ AND\ \{x2 \leq (x1+w1)\}\ AND\ \{(x2+w2) \geq (x1+w1)\}\ AND\ [\{(y2+h2) \leq (y1+h1)\}\ AND\ \{(y2+h2) \geq y1\}\ OR\ \{y2 \geq y1\ AND\ y2 \leq y1+h1\}]$
D	1:	$(x2+w2) \leq x1$
	2:	$x2 \geq (x1+w1)$
	3:	$y2 \geq (y1+h1)$
	4:	$(y2+h2) \leq y1$

Given such categories we will identify in a frame what the spatial relation is between two identified and categorized objects, this identification being in the pattern we adopted for the bbś $< x,y,w,h >$. LTN will receive this pair of bbś and as output, we will get the categorization of this spatial relation.

We then carried out the main experiment of this research with soccer videos. The first step with the classification and selection of frames by YOLO, followed by the reasoning of spatial relationships by LTN. To, finally, have the inference of the actions that occurred in the analyzed video.

5.1 Learning Groundings from Data

As described in this thesis, our classification occurs in each selected frame, so in this section we present the grounding for the characteristics needed in our scope.

The context of the characterization of our constants are the bounding boxes, bb, represented as follows:

$< Class, x, y, w, h >$
Class – Yolo classification
bb1: $< Class1, x1, y1, w1, h1 >$
bb2: $< Class2, x2, y2, w2, h2 >$

The perception of each spatial relation is defined by its coordinates identified by YOLO, 5. With this knowledge, we can define the grounding of our constraints, which make learning through LTN different from the others, as described above.

The YOLO Framework analyzes an image according to the general context of the image through the knowledge obtained in its pre-training. Therefore, this framework has the ability to perform such object prediction by its neural network. This makes it different from other methods like R-CNN and Fast R-CNN that analyze the image piece by piece.

This way we have the first part of our algorithm, which receives a video to separate all its frames, then selects from a range of time-spaced images, and finally recognizes the position of the object coordinates and classifies them according to the learning of the *preformed* method that makes up YOLO.

As discussed earlier, the output of YOLO will be the input to LTN, so that the spatial relations between pairs of bounding boxes will be identified through neural symbolic reasoning provided by this framework.

Using YOLO, each frame is represented as follows:

DOG 98% [525, 793, 500, 386]
PERSON 95% [1104, 58, 496, 1085]

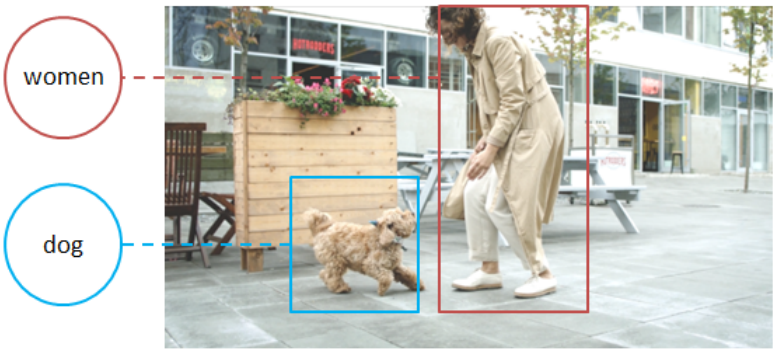


Figure 5.1: Bounding box identification by YOLO

Extracting the data present in the video using YOLO would be something represented in figure 5.1. The framework output identifies in each new frame the objects in a tuple formed by: object category; percentage associated with the category; bounding box coordinates.

When the LTN framework receives the file with the coordinates and categorization of the bounding boxes, it will instantiate each piece of information received according to the logic of its own algorithm.

In this case, the instances will be divided into constants, variables, class, and axioms. The constants will be the bounding boxes because through them we will carry out the reasoning for defining the variables of the spatial relations between their peers. These variables will be trained according to the axioms given to the context of this learning 3.3.4 to define the class of spatial relations.

So, for the Toy Example we use in this dissertation, we can have axioms that cause the grounding, 3.3, to recognize the dog toward the woman, the dog jumping, and the woman holding the dog, 5.2.

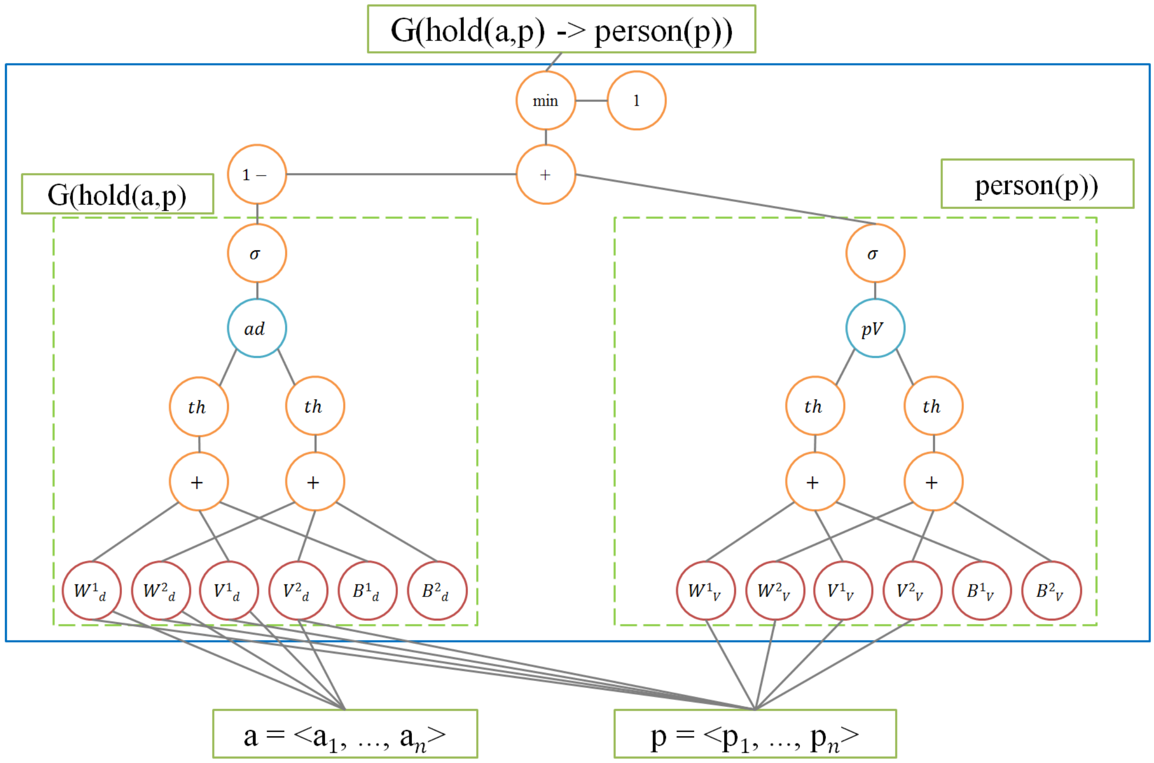


Figure 5.2: Example of Grounding in LTN for Toy Example

Reinforcing that such axioms are based on the positions of the coordinates of the bounding boxes, we thus have an adaptation of Cohen's definitions as described in 3.5.1.

These steps permit the possibility to infer the action from the video through the activity graph. To do this we sequentially join the frames that had the same classification, grouping constants that had the same spatial relation identification, this is shown in 5.3.

As described in 4.5 these clusters will be the nodes of our activity graph, and the edges of this graph will be the temporal relations based on Allen's temporal algebra 3.1.

This graph will have the edges indicating the next action, as we are dealing with isolated cases of actions between pairs of objects. Chapter 6 will describe the importance and symbolic power of this representation in complex actions.

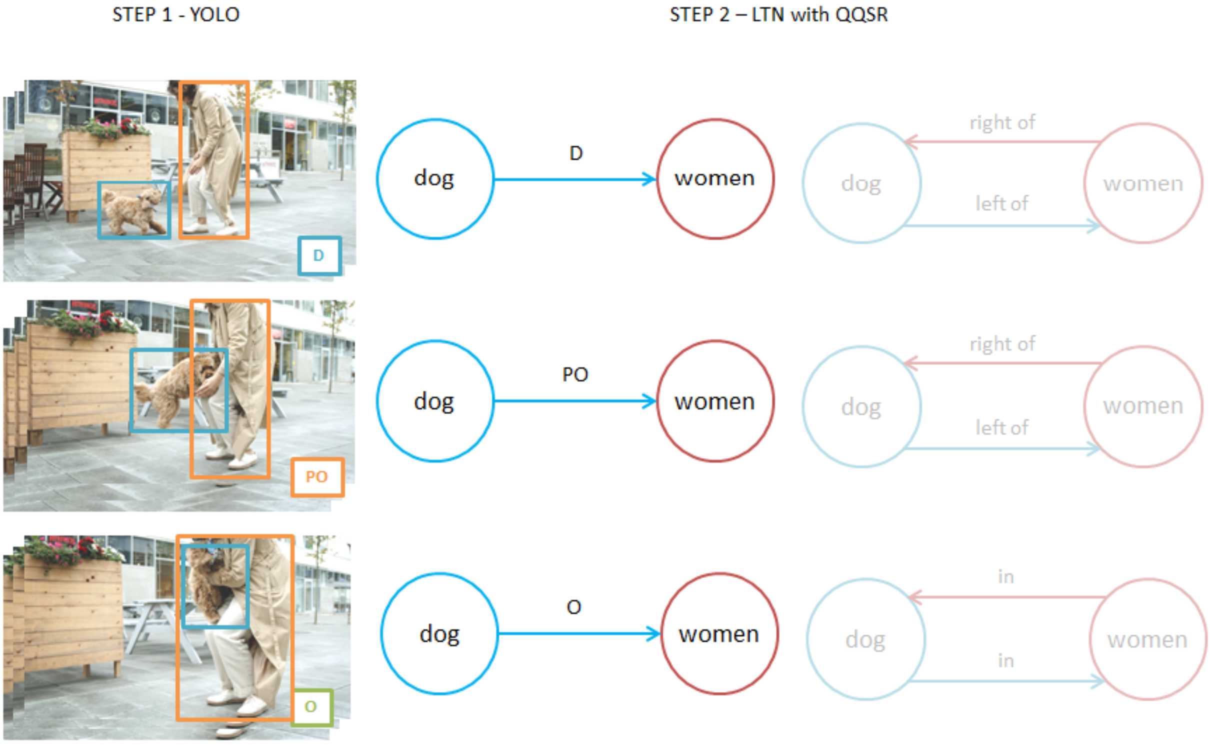


Figure 5.3: Steps of Construction of Activity Graph of Toy Example

Figure 5.3 shows the schema that represents what has been described so far divided by the steps of the discussed algorithm. The first step is the YOLO process, identifying the objects and classifying them.

5.2 Learning in Specific Context

The last step is the Activity Graph representation linking all the steps through Allen's algebra, as shown in 5.5. The next step guarantees the reasoning of spatial relations through prior knowledge of the LTN together with axioms based on Cohen's QQSTR relations.

This activity graph shows that the sequence of spatial relations represented is related to the dog's jump being followed by the woman's hug. Thus, we infer that the action of the video is to have object 1 being *"hold"* by object 2. Interestingly, because we treat relations in

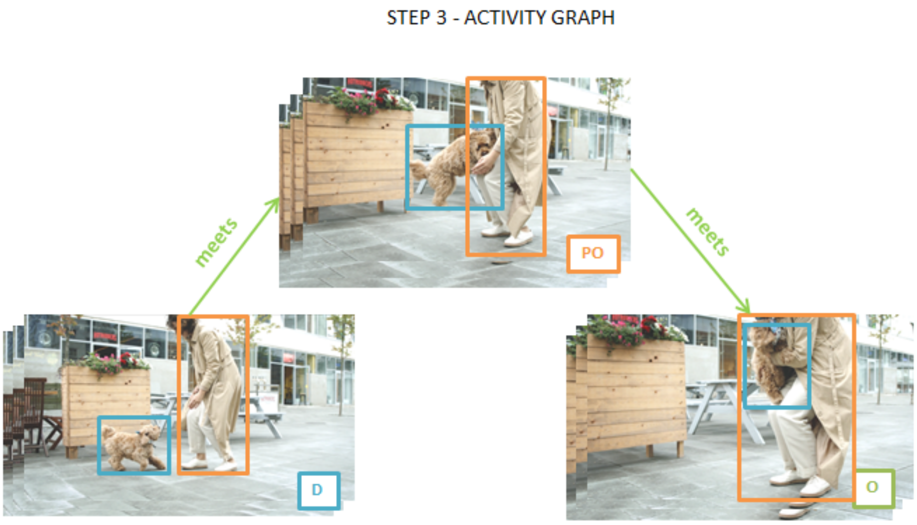


Figure 5.4: Steps of Construction of Activity Graph of Toy Example

an abstract and symbolic way, other videos can receive the exact inference, as the examples presented in 5.5.

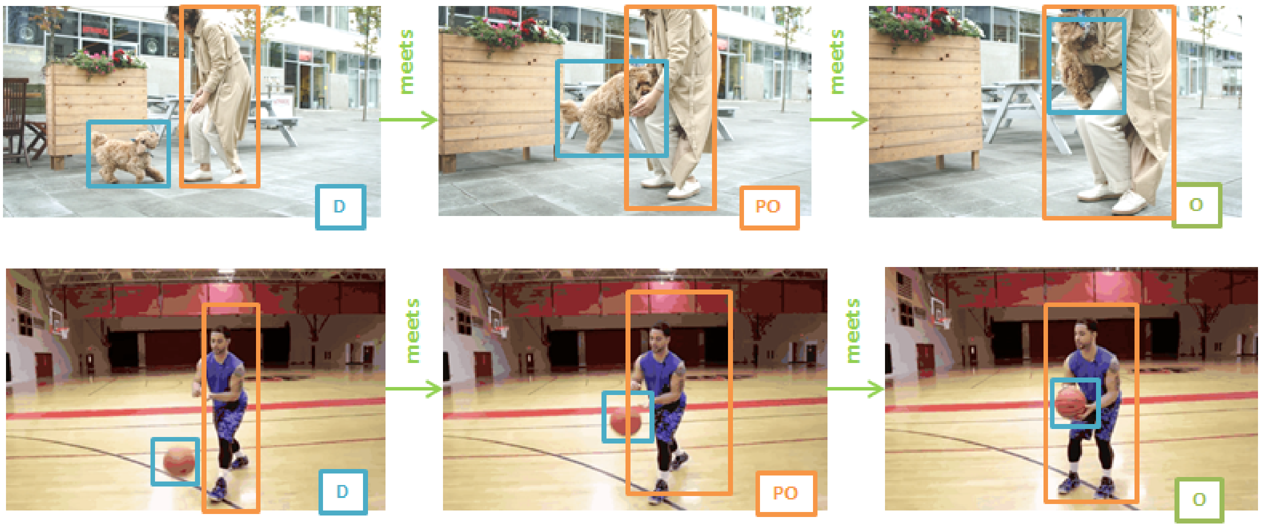


Figure 5.5: In both images we can identify the same action

We can observe that the sequence of instances of spatial relations follow a pattern which allows us to infer that in both videos the same action occurs.

5.2.1 Contextualizing for Soccer

By exploring a closed world of knowledge and definitions, we can obtain a totally specific approach to such a scenario. Thus, we will explore the algorithm discussed in this dissertation to identify specific spatial relations in soccer and the inferences of the actions they can represent.

Therefore, we will explore the construction of the axioms for this closed world of soccer. In this approach, as we always deal with pairs of objects, we chose to make the relations between players and relations between player and ball. In this way, reasoning about objects remains abstract for the spatial relations already presented: O *overlapping*, PO partial overlap and D *discrete*.

We divide theses spatial relations to be inferred in this approach into two groups. Because we always treat relations in pairs of objects, the first group for relations between two players and the second group for relations between player and ball:

Players

1. The tackling action 5.6 would be when two players from opposite teams meet in the same physical space, performing a kind of confrontation between them in order to make the opponent's locomotion difficult.

   $\forall xy(Player(x) \wedge Tackling(x,y) \rightarrow Player(y))$

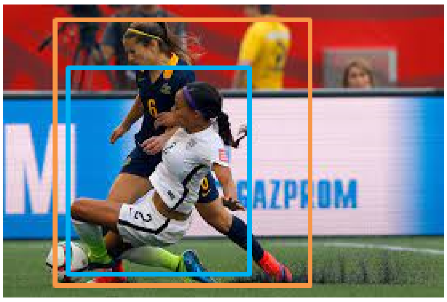


Figure 5.6: Tackling action has as final state the overlapping spatial relation between the players $O(player1, player2)$

2. The approaching action 5.7 happens when another player is so close to his opponent, but still there is no overlapping of them, in such a position it is possible to identify that the physical space of the players are getting closer with time.

   $\forall xy(Player(x) \wedge Approaching(x,y) \rightarrow Player(y))$

3. The Probing action 5.8 occurs when a player begins the process of approaching his opponent. In such conditions it is possible to notice the physical distance between them having the possibility of approximation with the evolution of time.

   $\forall xy(Player(x) \wedge ProbingPlayer(x,y) \rightarrow Player(y))$

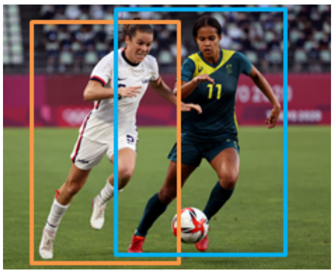


Figure 5.7: Approaching action has as final state the overlapping spatial relation $PO(player1, player2)$

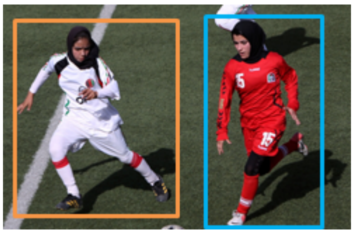


Figure 5.8: Probing action has as final state the discrete separation between the players $D(player1, player2)$

Ball Player

1. Controlling 5.9 is an action that occurs when the player is in possession of the ball in play. We identify it when both are physically very close, so their object bounding boxes are totally overlapped. In this moment the player in possession has dominion and is totally controlling the ball.

   $\forall xy(Player(x) \wedge Controlling(x,y) \rightarrow Ball(y))$

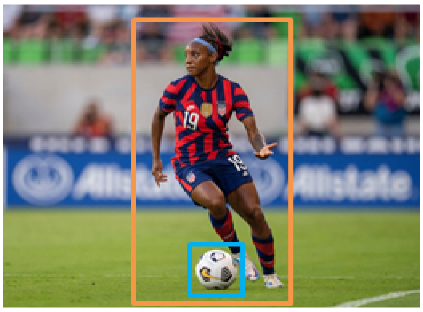


Figure 5.9: Controlling action has as final state the overlapping spatial relation between the player and the ball $O(player, ball)$

2. Pass 5.10 is the action identified when the player passes the ball to another, or when the player is receiving the ball. The inference of what actually happened depends on the spatial relations evidenced in the other frames of the video. In our context for this inference we only consider the physical distance between

the player and the ball and if this distance is not discrete and not completely overlapping we consider the partially overlapping spatial relation.

$\forall xy(Player(x) \wedge Passing(x,y) \rightarrow Ball(y))$

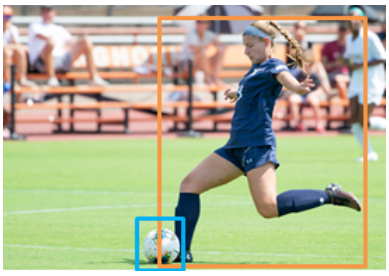


Figure 5.10: Passing action has as final state the partial overlapping spatial relation between the player and the ball $PO(player, ball)$

3. Losing 5.11 is the action when the player presents a physically noticeable distance from the ball, making it difficult for him to possess it, and thus we consider that there is no control relation between the player and the ball, inferring the action of losing the ball.

   $\forall xy(Player(x) \wedge Losing(x,y) \rightarrow Ball(y))$

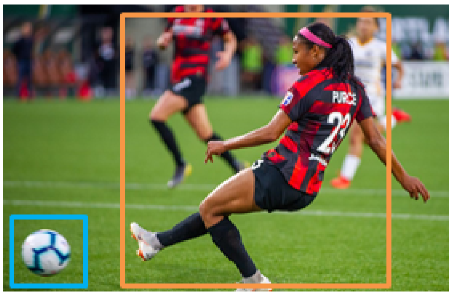


Figure 5.11: Losing action has as final state the discrete separation between the player and the ball of the spatial relation $D(player, ball)$

5.3 Applying the algorithm to short videos about soccer

We consider the evolution of each action described in the previous section, 5.2.1, in our proposed algorithm 4.4.1. In 5.12 we present each step of the algorithm specifying the knowledge for soccer and in the following list we have the final result of each action.

The definitions of spatial relations in O, PO or D occur from Cohen's QQSTR approach already demonstrated in 3.11. This identification occurs by analyzing the coordinates of the bounding boxes over their behavior *qualitatively* as discussed in 4.2.1.

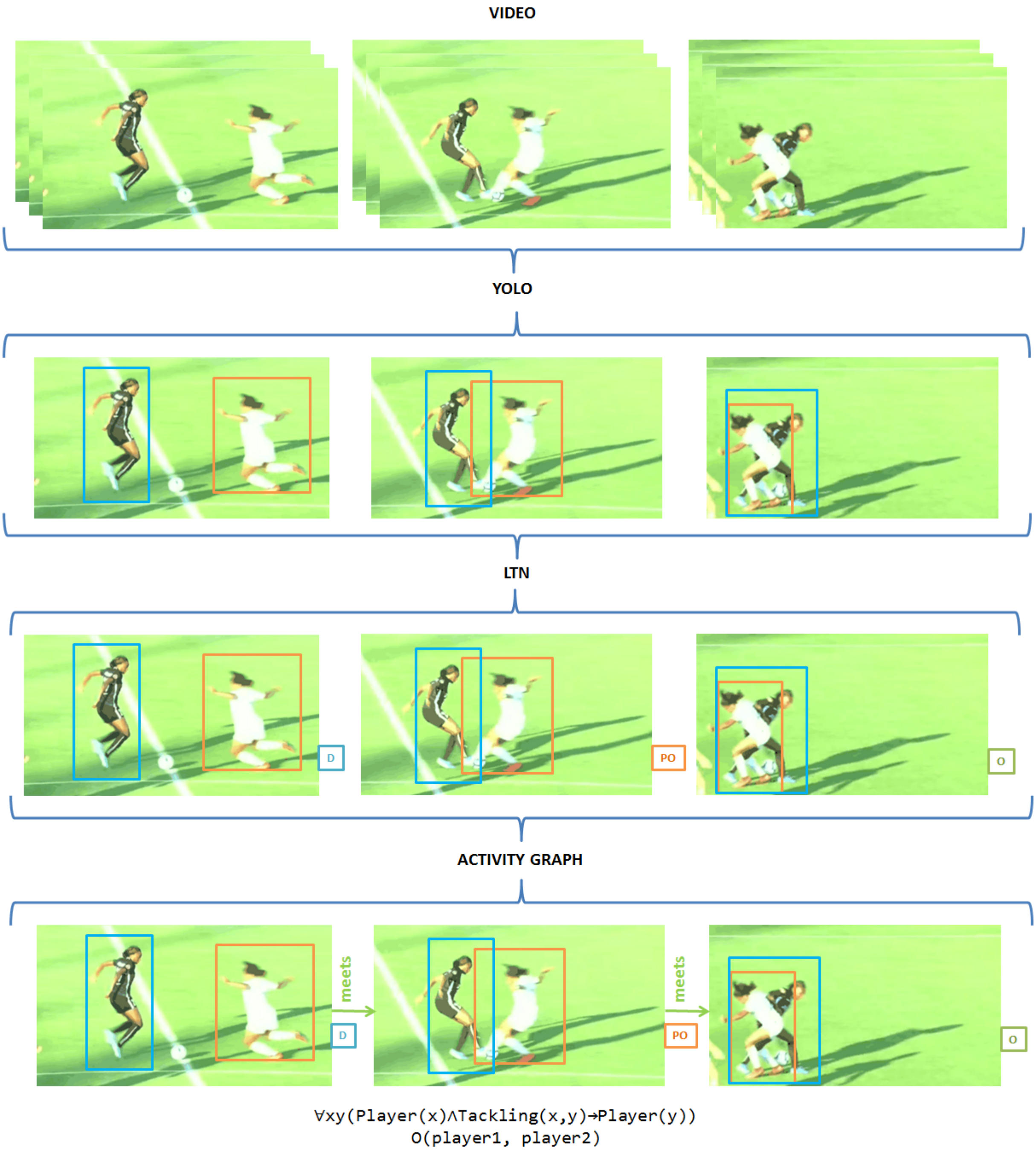


Figure 5.12: Steps of the proposed algorithm for action inference Probing during a soccer match

- Tackling/Controlling O
  $\forall xy(Player(x) \land O(x,y) \rightarrow Tackling(x,y) \lor Controlling(x,y))$

- Approaching/Passing PO
  $\forall xy(Player(x) \wedge PO(x,y) \rightarrow Approaching(x,y) \vee Passing(x,y))$

- Probing Player/Losing D
  $\forall xy(Player(x) \wedge D(x,y) \rightarrow ProbingPlayer(x,y) \vee Losing(x,y))$

The spatial relations mentioned have similar characteristics, for instance, we can describe that the spatial relation related to the Tackling and Controlling actions is O, in which the related objects have the distance of overlap between them.

For the soccer context where the relations between objects are partially overlapping, we have the actions related to Approaching and Passing. And finally, the discrete relations between objects are associated with the actions Probing Player and Losing.

5.4 Instantiating the aspects in LTN

As mentioned at the beginning of this subsection we have defined only two types of possible object identifications in this environment: player and ball, just as the relations that can be inferred are limited to: *Tackling, Approaching, Probing* for player relations, and *Controlling, Passing, and Losing* for player-ball relations.

In LTN we declare spatial relations as classes, see below:

class$_P = ltn.Constant(0, trainable = False)$

class$_{PO} = ltn.Constant(1, trainable = False)$

class$_D = ltn.Constant(2, trainable = False)$

These classes will be managed with the constants to identify the variable classifications that are the spatial relations between the pairs of bounding boxes.

x$_p = ltn.Variable(”x_p”, features[label\_position == P])$

x$_{po} = ltn.Variable(”x_{po}”, features[label\_position == PO])$

x$_d = ltn.Variable(”x_d”, features[label\_position == D])$

Below we demonstrate each activity graph of the following actions that we discussed: Approaching 5.13, Probing 5.15, Controlling 5.15, Passing 5.16 and Losing 5.11.

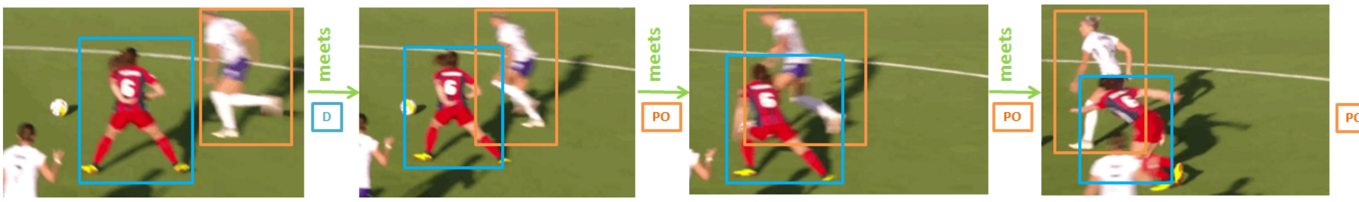


Figure 5.13: Approaching Activity Graph

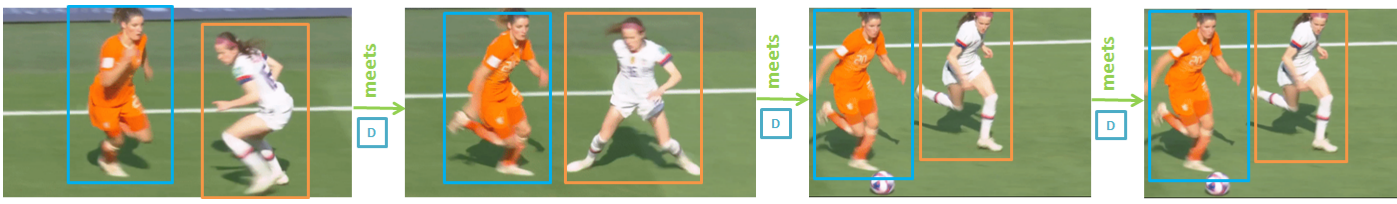


Figure 5.14: Probing Activity Graph

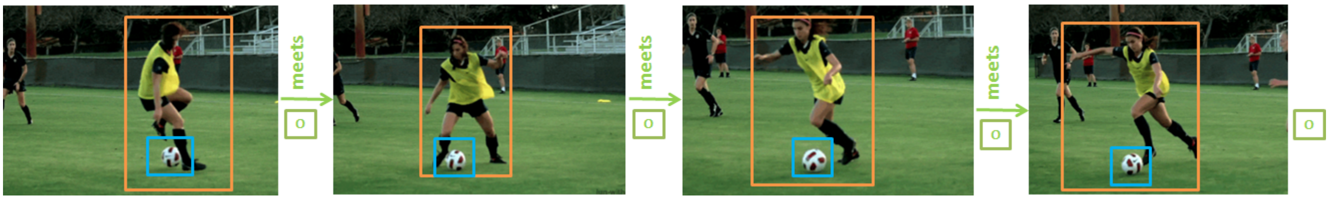


Figure 5.15: Controlling Activity Graph

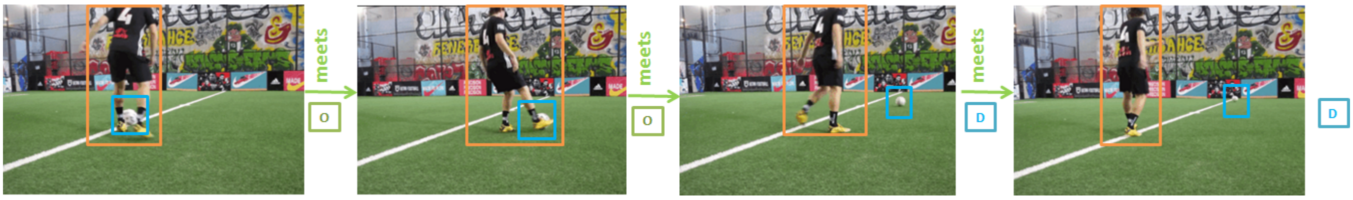


Figure 5.16: Passing Activity Graph

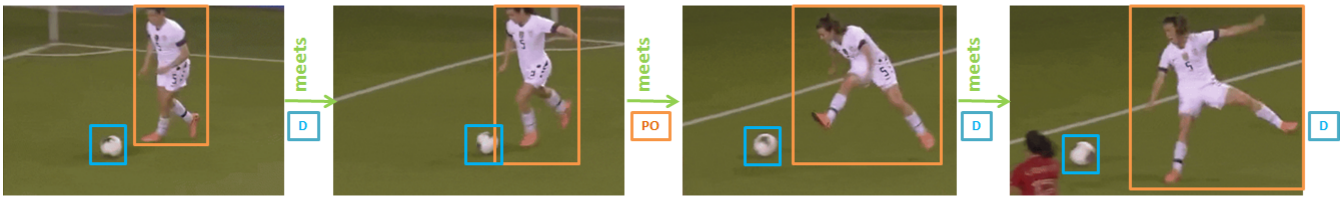


Figure 5.17: Losing Activity Graph

In this way, we identified the superiority of our hypothesis using the neural symbolic approach provided by the LTN in the verification and validation of the reasoning performed by the algorithm. Demonstrating the strong representation of the information by the symbolic part of the algorithm, for understanding the human being. As well as the reasoning performed by neural networks with symbolic constraints to improve the identification of spatial relations.

Chapter 6

Conclusion

Right from the very beginning, this project was based on the principle of studying the possibility of reasoning about videos through a neural symbolic approach, initially studying the state of the art of video analytics, what are the main challenges and the best approaches according to the different objectives proposed and discussed at the beginning of this dissertation.

The main objective was achieved through the divide and conquer approach, more specifically LTN, which started by the search for the smallest possible division to start our project. Thus we began the search for a method that would better decode the spatial characteristics of videos.

In this way, we identified the need to isolate the characteristics of a video to perform its classification. For this reason, we chose QQSTR as a method for a detailed approach to the physical positions of objects in the images, identifying that it would not be necessary to carry out a quantitative analysis of the information available for analysis, such as R-CNN.

This approach allows the use of the LTN framework to analyze the spatial relations that occur during the video. However, LTN analyzes information statically, and for this reason, we chose the YOLO framework to prepare the data to be reasoned by the LTN.

YOLO on the other hand receives a video and separates it into several frameworks, selecting frames at each time interval to optimize the video processing. After the frames are selected, the detection of objects in the images is performed and identified by bounding boxes.

These bounding boxes, in turn, are categorized by the same YOLO framework for recognizing the possible objects detected, based on a pre-trained knowledge of YOLO itself.

Only after this process is the LTN able to reason about the spatial relations between pairs of previously detected and categorized objects. This reasoning takes place from the neuro-symbolic domain of LTN, which through prior knowledge of positive and negative examples of classifications further improves knowledge of LTN along with axioms of first-order logic.

The proposal of this research was presented at the Black in AI workshop, at the 2019 NeurIPS congress, in the article "Learning Spatial Relations LTN with Qualitative and Quantitative Representations of Spatial-Temporal Resources". The proposal of integration of the time restrictions in the LTN was presented so that its applications of static environments represented in images, could be extended in applications in dynamic environments, as data present in the video.

The article achieved such visibility by demonstrating that the neural symbolic approach unites the strong data representation of symbolic methods with the current advances in artificial intelligence achieved by neural networks. As a combination of methods with goals and ascension in different AI eras, neural symbolism brings advantages to both methods.

Thus there were challenges in understanding the union of these methods and their relations in this new approach, comparing and adapting the methods related to the research were moments that required more dedication to better understand the scope of the study and the influences that the differences could cause.

We found that by using the symbolic neural network during the video action identification algorithm, it was possible to obtain information from the entire analysis process to complete the final identification because we treat data in a symbolic way. This allowed the manipulation of these data and their description in the activity graph.

Through this study, it was possible to verify that by using symbolic neural networks, the proposed video action identification algorithm allows obtaining information and treating them in a symbolic way, having information about the entire analysis process for the conclusion of the identification of action in a video.

As shown in schema 5.5 it is possible to use the hypothesis addressed in this research for different contexts of videos and related objects. For a more specific description, we detail for the desired context, as we have presented for soccer.

This research aimed to demonstrate the efficiency of the method presented and its advantage over other algorithms due to its strong representation through a study applied in a context of soccer videos.

This is a fundamental work for future research of more complex actions and videos. It is possible to perform abstraction for several approaches to data that are not only visual but that progress in time and space. Supporting that it is possible to obtain more representation of neural methods by unifying with a symbolic approach, making the black box of traditional neural methods clearer.

Based on the information obtained, our objective is to work on expanding this approach to complex actions and longer videos.

This will be possible by representing more actions on the same activity graph, using other temporal relations on the directed edges of the graph, and demonstrating the relations between more objects.

In the context of a soccer game, it would be possible to analyze three objects, for example, we could identify the actions that take place in a video in which there are two opposing players, only one with possession of the ball while the other tries to get it as represented in figure 6.1.

For such a problem we would analyze it in the same way, between pairs of objects. So we would have the relation between the ball and Player 1, the ball and Player 2, and finally the spatial relation between the players. Each instance of these relations would have a temporal relation according to what occurs in the video. Adding new temporal relations such as starts on, ends on, and during, for example.

Notice that there are several spatial relations in each node of the graph represented in 6.1, this way we could also make a more refined analysis of the temporal relations in the nodes themselves, see the figure 6.2.

As relates to the main objective of this research, it was possible to develop the proposed algorithm. Our approach represents the data from the information obtained from the video and manipulates it through symbolic AI that is easily interpreted by humans. Making our method more interpretable than others already mentioned.

For the partial goals, YOLO was the framework chosen to identify objects and their positions in the video due to its superiority compared to other methods. This is because the analysis is performed for the entire image and not for pieces, as is the case of R-CNN. It was possible to provide a description of the spatial relations in symbolic neuro-networks by mapping the proposed relationships for LTN, providing the representation of spatial relations by adapting the LTN relations based on the QQSTR method as O overlaps, PO partially overlaps and D discretely does not overlap. More so, the activity graph allows one to model the information identified to infer the action that took place in the video.

Through the initiative of this research, it was possible to initiate a partnership with the Faculdade de Educação Física e Fisioterapia (FEFF) of UFAM, thereby providing tactical and teaching knowledge to the students and teachers of the project. The project consists of collecting data from soccer games, fundamental movements, and player progression during the training sessions carried out by the project. It is also possible to generate a database for future works cited, increase FEFF's technology resources, and the emergence of new project partnerships.

Through this research, we seek to expand the possibilities of using symbolic neural networks and contribute to the field of video analytics by integrating knowledge from various

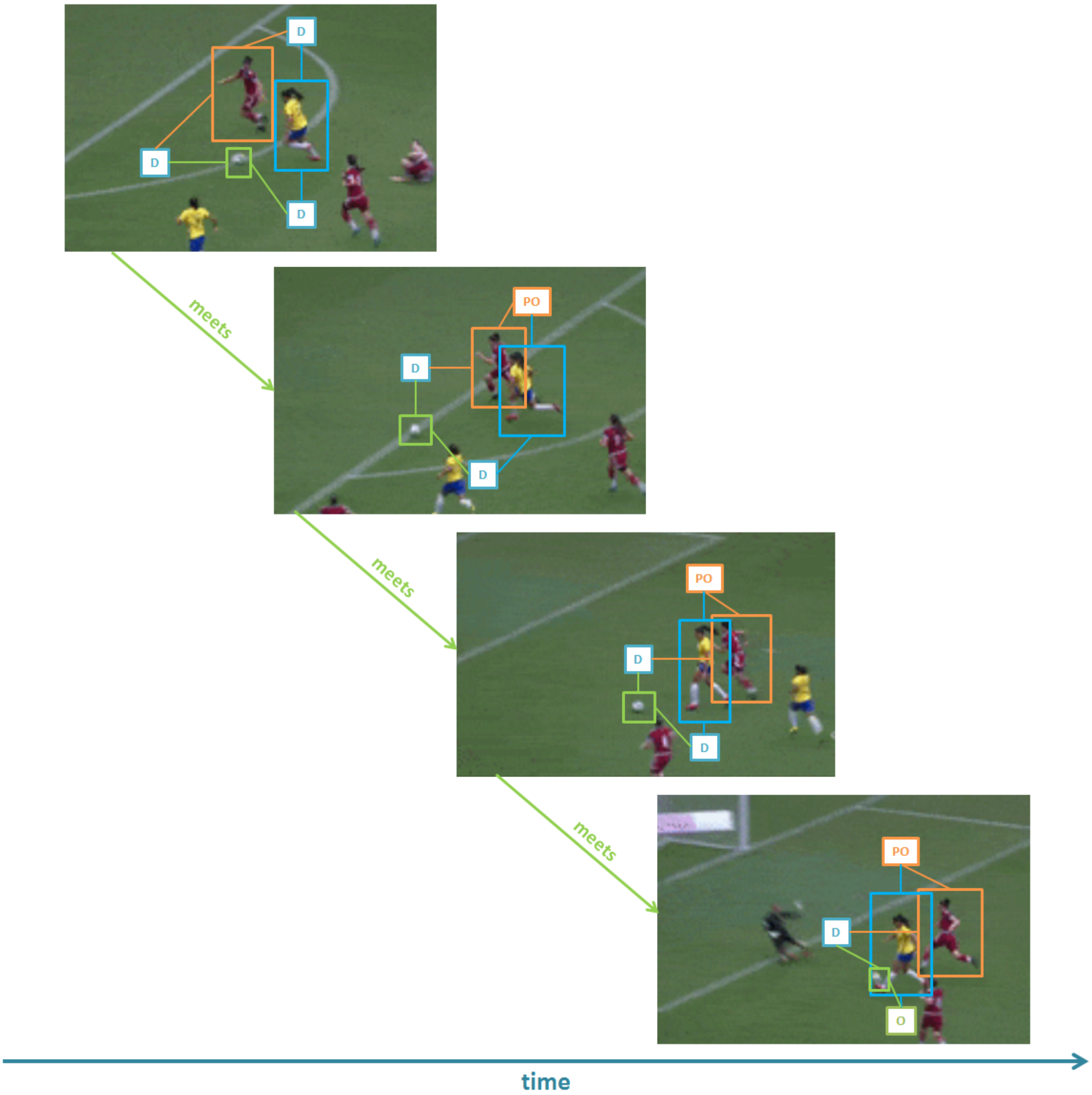


Figure 6.1: Hypothesis of our approach in a complex context

segments of AI, video processing and analytics, and syntactic and semantic approaches to the data used. Furthermore, by advancing the analysis to a specific context of several objects interacting with each other in the same space of time, such advancement can also be expanded to other contexts of progressive data.

The research contributions lie in the investigation provided for the use of neuro-symbolic methods in rationalizing data as symbolic data.

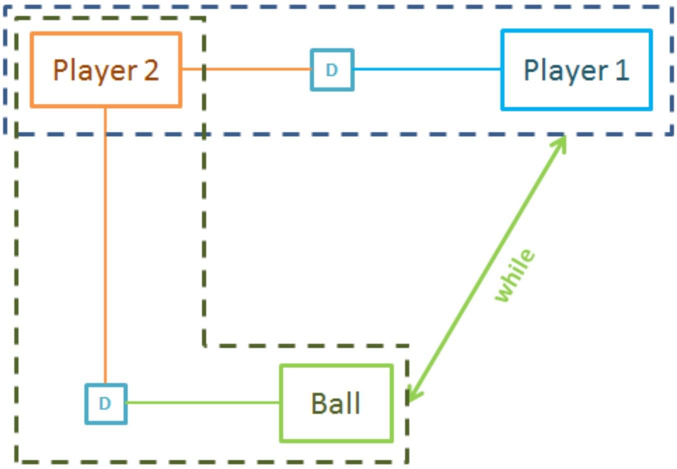


Figure 6.2: Hypothesis of our approach in a complex context about nodes of the activity graph

Appendix A

Syntax of spatial relationships

This chapter provides formal definitions that will be used to verify some properties of the model. As they are a work under development, they are put in an appendix. This work extends LTN spatial relationships between pairs of bounding boxes:

Table regarding definitions of spatial relationships.

Table A.1: New spatial relationships

$\mathscr{G}(inSideLeftAltAbove(b,b^{'}))$ $\mathscr{G}(inAltAbove(b,b^{'}))$ $\mathscr{G}(inSideRightAltAbove(b,b^{'}))$ $\mathscr{G}(inSideRight(b,b^{'}))$ $\mathscr{G}(inSideRightAltBelow(b,b^{'}))$ $\mathscr{G}(inAltBelow(b,b^{'}))$ $\mathscr{G}(inSideLeftAltBelow(b,b^{'}))$ $\mathscr{G}(inSideLeft(b,b^{'}))$	$x_0(b) \geqslant x_0(b^{'})$ $y_0(b) \leq y_0(b^{'})$ $x_1(b) \leq x_1(b^{'})$ $y_1(b) \geqslant y_1(b^{'})$ $regardlessofangle$
$\mathscr{G}(edgeAltAboveSideLeft(b,b^{'}))$	$x_0(b) < x_0(b^{'})$ $x_1(b) < x_1(b^{'})$ $y_0(b) > y_0(b^{'})$ $y_1(b) < y_1(b^{'})$ 135^o
$\mathscr{G}(edgeAltAbove(b,b^{'}))$	$x_0(b) > x_0(b^{'})$ $x_1(b) < x_1(b^{'})$ $y_0(b) > y_0(b^{'})$ $y_1(b) < y_1(b^{'})$ 90^o

$\mathscr{G}(edgeAltAboveSideRight(b,b^{'}))$	$x_0(b) > x_0(b^{'})$ $x_1(b) > x_1(b^{'})$ $y_0(b) > y_0(b^{'})$ $y_1(b) < y_1(b^{'})$ 45^o
$\mathscr{G}(edgeSideRight(b,b^{'}))$	$x_0(b) > x_0(b^{'})$ $x_1(b) > x_1(b^{'})$ $y_0(b) < y_0(b^{'})$ $y_1(b) > y_1(b^{'})$ $0^o or 360^o$
$\mathscr{G}(edgeAltAboveSideRight(b,b^{'}))$	$x_0(b) > x_0(b^{'})$ $x_1(b) > x_1(b^{'})$ $y_0(b) < y_0(b^{'})$ $y_1(b) < y_1(b^{'})$ 315^o
$\mathscr{G}(edgeAltAbove(b,b^{'}))$	$x_0(b) > x_0(b^{'})$ $x_1(b) < x_1(b^{'})$ $y_0(b) < y_0(b^{'})$ $y_1(b) < y_1(b^{'})$ 270^o
$\mathscr{G}(edgeAltAboveSideLeft(b,b^{'}))$	$x_0(b) < x_0(b^{'})$ $x_1(b) < x_1(b^{'})$ $y_0(b) < y_0(b^{'})$ $y_1(b) < y_1(b^{'})$ 225^o
$\mathscr{G}(edgeSideLeft(b,b^{'}))$	$x_0(b) < x_0(b^{'})$ $x_1(b) < x_1(b^{'})$ $y_0(b) < y_0(b^{'})$ $y_1(b) > y_1(b^{'})$ 180^o
$\mathscr{G}(outAltAboveSideRight(b,b^{'}))$ $\mathscr{G}(outAltAboveSide(b,b^{'}))$ $\mathscr{G}(outAltAboveSideLeft(b,b^{'}))$	$y_0(b) > y_0(b^{'})$ $y_1(b) > y_1(b^{'})$
$\mathscr{G}(outSideRight(b,b))$	$x_0(b) < x_0(b^{'})$ $x_1(b) < x_1(b^{'})$
$\mathscr{G}(outSideLeft(b,b^{'}))$	$x_0(b) > x_1(b^{'})$

$$\begin{array}{ll}
\mathscr{G}(outAltBelowSideRight(b,b^{'})) & \\
\mathscr{G}(outAltBelowSide(b,b^{'})) & y_0(b) < y_1(b^{'}) \\
\mathscr{G}(outAltBelowSideLeft(b,b^{'})) &
\end{array}$$

Bibliography

[1] R. Borges, *A neural-symbolic system for temporal reasoning with application to model verification and learning*. PhD thesis, City University London, 2012.

[2] I. Donadello, L. Serafini, and A. S. d'Avila Garcez, "Logic tensor networks for semantic image interpretation," in *IJCAI*, 2017.

[3] J. Redmon, "Yolov3: An incremental improvement/joseph redmon, ali farhadi-university of washington," 2018.

[4] A. G. Cohn, "Qualitative spatial representations," in *Proc. IJCAI-99 Workshop on Adaptative Spatial Representations of Dynamic Environments*, Citeseer, 1999.

[5] S. Mate, T. Bürkle, L. Kapsner, D. Toddenroth, M. Kampf, M. Sedlmayr, I. Castellanos, H.-U. Prokosch, and S. Kraus, "A method for the graphical modeling of relative temporal constraints (preprint)," 03 2019.

[6] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courville, and B. Schiele, "Movie description," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 94–120, 2017.

[7] N. Aafaq, A. Mian, W. Liu, S. Z. Gilani, and M. Shah, "Video description: A survey of methods, datasets, and evaluation metrics," *ACM Computing Surveys (CSUR)*, vol. 52, no. 6, pp. 1–37, 2019.

[8] D. Koller, N. Heinze, and H.-H. Nagel, "Algorithmic characterization of vehicle trajectories from image sequences by motion verbs," in *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 90–91, IEEE Computer Society, 1991.

[9] M. Brand, "The" inverse hollywood problem": From video to scripts and storyboards via causal analysis," in *AAAI/IAAI*, pp. 132–137, Citeseer, 1997.

[10] J. Tayyub, A. Tavanai, Y. Gatsoulis, A. G. Cohn, and D. C. Hogg, "Qualitative and quantitative spatio-temporal relations in daily living activity recognition," in *Asian Conference on Computer Vision*, pp. 115–130, Springer, 2014.

[11] S. Burkhardt, J. Brugger, N. Wagner, Z. Ahmadi, K. Kersting, and S. Kramer, "Rule extraction from binary neural networks with convolutional rules for model validation," *arXiv preprint arXiv:2012.08459*, 2020.

[12] L. Serafini and A. S. d. Garcez, “Learning and reasoning with logic tensor networks,” in *Conference of the Italian Association for Artificial Intelligence*, pp. 334–348, Springer, 2016.

[13] I. Donadello, *Semantic image interpretation-integration of numerical data and logical knowledge for cognitive vision*. PhD thesis, University of Trento, 2018.

[14] I. Tiddi *et al.*, “Neuro-symbolic architectures for context understanding,” *Knowledge Graphs for eXplainable Artificial Intelligence: Foundations, Applications and Challenges*, vol. 47, p. 143, 2020.

[15] M. R. Tenório, E. Mota, J. M. Howe, and A. Garcez, “Learning about actions and events in shared nemus,” in *CEUR Workshop Proceedings*, vol. 2003, CEUR Workshop Proceedings, 2017.

[16] A. d. Garcez and L. C. Lamb, “Neurosymbolic ai: The 3rd wave,” *arXiv preprint arXiv:2012.05876*, 2020.

[17] M. Sridhar, A. G. Cohn, and D. C. Hogg, “Learning functional object-categories from a relational spatio-temporal representation,” in *ECAI*, 2008.

[18] I. Donadello and L. Serafini, “Compensating supervision incompleteness with prior knowledge in semantic image interpretation,” in *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2019.

[19] A. S. d’Avila Garcez, “Neurons and symbols: a manifesto,” in *Dagstuhl Seminar Proceedings*, Schloss Dagstuhl-Leibniz-Zentrum fÃ1/4r Informatik, 2010.

[20] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, *et al.*, “Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai,” *Information Fusion*, vol. 58, pp. 82–115, 2020.

[21] S. Bader, *Neural-symbolic integration*. PhD thesis, Dresden University of Technology, 2009.

[22] A. S. Garcez, L. C. Lamb, and D. M. Gabbay, *Neural-symbolic cognitive reasoning*. Springer Science & Business Media, 2008.

[23] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, *et al.*, “Relational inductive biases, deep learning, and graph networks,” *arXiv preprint arXiv:1806.01261*, 2018.

[24] H. Cohen and C. Lefebvre, *Handbook of categorization in cognitive science*. Elsevier, 2005.

[25] M. K. Sarker, L. Zhou, A. Eberhart, and P. Hitzler, “Neuro-symbolic artificial intelligence: Current trends,” *arXiv preprint arXiv:2105.05330*, 2021.

[26] A. Bennetot, J.-L. Laurent, R. Chatila, and N. Díaz-Rodríguez, “Towards explainable neural-symbolic visual reasoning,” in *NeSy Workshop IJCAI*, 2019.

[27] S. Badreddine, A. d'Avila Garcez, L. Serafini, and M. Spranger, "Logic tensor networks," 2021.

[28] L. Serafini and A. d. Garcez, "Logic tensor networks: Deep learning and logical reasoning from data and knowledge," *arXiv preprint arXiv:1606.04422*, 2016.

[29] L. Serafini and M. Spranger, "A practical introduction to learning and reasoning with ltns." Neural-Symbolic Learning and Reasoning with Constraints Tutorial, 2018.

[30] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille, "Detect what you can: Detecting and representing objects using holistic models and body parts," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1971–1978, 2014.

[31] C. Fellbaum, "Wordnet: An electronic lexical database," *Language, Speech, and Communication. MIT Press, Cambrige, MA*, 1998.

[32] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *European conference on computer vision*, pp. 852–869, Springer, 2016.

[33] J. Redmon, "Darknet: Open source neural networks in c." http://pjreddie.com/darknet/, 2013–2016.

[34] J. Tayyub, A. Tavanai, Y. Gatsoulis, A. G. Cohn, and D. C. Hogg, "Qualitative and quantitative spatio-temporal relations in daily living activity recognition," in *Asian Conference on Computer Vision*, pp. 115–130, Springer, 2014.

[35] B. Nebel and H.-J. Bürckert, "Reasoning about temporal relations: a maximal tractable subclass of allen's interval algebra," *Journal of the ACM (JACM)*, vol. 42, no. 1, pp. 43–66, 1995.

[36] J. Zhang, Y. Han, J. Jiang, Z. Zhou, D. An, J. Liu, and Z. Song, "A feature selection framework for video semantic recognition via integrated cross-media analysis and embedded learning," *EURASIP Journal on Image and Video Processing*, vol. 2019, no. 1, p. 44, 2019.

[37] S. Aakur, F. D. de Souza, and S. Sarkar, "Going deeper with semantics: Video activity interpretation using semantic contextualization," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 190–199, IEEE, 2019.

[38] M. Sridhar, A. G. Cohn, and D. C. Hogg, "Learning functional object categories from a relational spatio-temporal representation," in *ECAI 2008: 18th European Conference on Artificial Intelligence (Frontiers in Artificial Intelligence and Applications)*, pp. 606–610, IOS Press, 2008.