University of Padova

Department of Mathematics

Master Thesis in Data Science

Towards Explainability in Knowledge Enhanced Neural Networks

Supervisor
Luciano Serafini
University of Padova

Co-supervisor
Alessandro Daniele
Fondazione Bruno Kessler

Master Candidate
Riccardo Mazzieri

Academic Year
2020-2021

Dedication.

Abstract

Research on Deep Learning has achieved remarkable results in recent years, mainly thanks to the computing power of modern computers and the increasing availability of large data sets. However, deep neural models are universally considered black boxes: they employ sub-symbolic representations of knowledge, which are inherently opaque to human beings trying to derive explanations. In this work, we first give a survey on the research field of Explainable AI, providing more rigorous definitions of the concepts of interpretability and explainability. We then delve deeper in the research field of Neural Symbolic Integration, which tackles the task of integrating the statistical learning power of machine learning with the symbolic and abstract world of logic. Specifically, we analyze Knowledge Enhanced Neural Networks (KENN) [1], a special kind of residual layer for neural architectures which makes it possible to inject symbolic logical knowledge inside a neural network. We describe and analyze experimental results on relational data on the task of collective classification, and study how KENN is able to automatically learn the importance of logical rules from the training data. We finally review explainability methods for KENN, proposing ways to extract explanations for the predictions provided by the model.

Contents

Listing of figures

Listing of tables

Listing of acronyms

ML . . . . . Machine Learning

XAI . . . . . Explainable Artificial Intelligence

NN . . . . . Neural Network

KENN . . . . . Knowledge Enhanced Neural Networks

TBF . . . . . t-conorm Boost Function

CE . . . . . Clause Enhancer

KE . . . . . Knowledge Enhancer

SBR . . . . . Semantic Based Regularization

LTN . . . . . Logic Tensor Networks

RNM . . . . . Relational Neural Machines

NeSy . . . . . Neural Symbolic Integration

1 Introduction

Deep Learning (DL) research has been moving faster and faster in recent years. Deep models continue to deliver state-of-the-art results in a wide variety of AI applications, large companies are investing huge amounts of money on DL research, the general reputation of deep learning continues to grow, and the enthusiasm towards data-driven AI only seems likely to increase with time. However, the limitations of this approach have already been identified and criticized [4]: in fact, while the remarkable results achieved by deep neural models are undeniable, their limitations are often underestimated or ignored.

The first critical aspect of DL is its exclusive dependence on data: this learning framework has proven to be very effective for many specific tasks; however, without enough data, DL models still lack the ability to learn even basic concepts, due to their inability to reason in symbolic and abstract terms. For example, if you verbally describe a unicorn to a child, he or she will likely be able to recognize it on a television program, to explain to others what a unicorn is, and perhaps even to produce a representation in a drawing of what he or she has learned from the description. This is because the concept of a unicorn is simple to elaborate for a human, being a simple composition of already known objects. Specifically, this composition can be easily expressed with logical formulas, which do not require any kind of data to be able to express new information. In contrast, to teach an NN what a unicorn is, one would have to provide it with hundreds or thousands of pictures, each with a label that tells it whether that is a unicorn or not. This approach looks unnecessary for such a trivial task; however, DL models have not yet been fully integrated with symbolical reasoning and prior knowledge in order to solve this

kind of task as a human would.

Another critical aspects of DL models is their black-box nature: deep NNs have millions or even billions of parameters that can't be directly examined and interpreted by humans. While model transparency is not always necessary, this opaqueness can be dangerous in some critical domains such as medical and financial ones, and poses the risk of heavily biased models, as pointed out in [5].

These, and many other critical aspects of DL motivate this work of thesis, which will be focused around two specific active research fields. The first is Explainable AI (XAI), where the focus is on the task of deriving explanations from ML systems with the aim of achieving more transparent models, boosting trust towards data-driven AI as a result. We will then delve deeper into the field of Neural Symbolic Integration (NeSy), which tackles the challenge of integrating logical knowledge inside NNs. Specifically, this thesis focuses on Knowledge Enhanced Neural Networks (KENN) [1], a special kind of NN layer designed to inject logical knowledge inside a neural model, in order to improve its predictive capabilities. We will report theoretical details, as well as experimental results, together with an analysis of the explainability of KENN.

This work of thesis was conducted during my internship at the Bruno Kessler Foundation in Trento, Italy. During this experience, I worked alongside researcher Alessandro Daniele and professor Luciano Serafini on experiments and further research on KENN, which led to the development of this work.

This thesis is organized as follows. In Chapter 2 we will discuss in depth the concept of explainability in Machine Learning and report the last developments in the field of XAI. In Chapter 3 we will focus on KENN. Specifically, in Section 3.1 we'll describe the theoretical framework behind the model. In Section 3.2 we'll see its architecture, and in Section 3.5 we will report results of the experiments performed on relational data, on a collective classification task. Finally, in Section 3.6 will discuss about the explainability of KENN and devise ways to derive explanations from it.

2

Explainability in Machine Learning

In recent years, research on Deep Learning (DL) has achieved remarkable results in a wide variety of domains, going from image recognition, speech recognition, machine translation, playing games, and many others, in some cases even outperforming human capabilities [6, 7, 8]. Although DL is not a new research topic [9, 10], its popularity and capabilities skyrocketed only in the last decade: this has been possible mostly thanks to the always growing availability of new data, together with the increase in computational power of modern machines. Despite the quick and huge success, researchers in the field have already shown some perplexities and moved some critiques towards this approach [4, 11]: is DL really the future of Artificial Intelligence (AI)? Will Neural Networks (NN) be able to give a good approximation of the human brain? Leaving aside such overwhelming questions, we should still be interested in what DL is capable to do at the moment, and what capabilities it still lacks. One particular aspect for which DL has been criticized is its lack of transparency: in fact, deep models have millions or even billions of learnable parameters, which are not characterizable in any human interpretable way. This results in opaque molearnable e no human supervisor can, ultimately, inteany t what the model has learned by simply inspecting its internal structure. To refer to this undesirable property of deep NNs, the term "black box" is commonly used. DL models also present other critical and perplexing issues: in [12] the authors made a neural network misclassify an image by applying an hardly perceptible perturbation, found by maximizing the network's prediction error. Such adversarial examples have also been found to be somewhat universal and not just the result of overfitting [13]. Similarly, authors in [14] show how deep neural networks are easily

fooled into misclassifying inputs with no resemblance to their true category. This kind of issues pose serious doubts about the ability of NN to learn general representations: indeed, if such networks can generalize well, how can they be confused by what we see as nearly indistinguishable images? Adversarial examples are not confined to the field of computer vision: natural language networks can also be fooled as shown in [15, 16]. Furthermore, it has been found that in several applications, DL models present strong biasedness. One example is reported in [17], where the authors show how word embeddings trained on Google News articles exhibit strong female/male gender stereotypes due to biases in the training data. Susceptibility to unintuitive errors remains therefore a pervasive problem in DL and no robust solution has been found for them so far. Such issues contribute to generate mistrust, and threaten to slow down or even hinder the prevailance of AI in some applications, due to the high potential of unexpected behavior and lack of verifiability of solutions. In light of such problems, Explainable Artificial Intelligence (XAI) has become an area of interest in the research community: it tackles the important problem that complex machines and algorithms often cannot provide insights into their behavior and thought processes. The need for XAI is now even more urgent: the renewed EU General Data Protection Regulation (GDPR) could require AI providers to provide users with explanations of the results of ML systems based on their personal data. This clearly affects the industry in a huge way: indeed, the GDPR may hinder or even prohibit the use of "black box" models which don't offer explanations for their decisions, when based on users' personal data (think for example to recommender systems). This is also referred to as the "right to explanation" [18]. The need for XAI has been also expressed by the statement on algorithmic transparency and accountability released by the Association for Computing Machinery [19], and by the XAI program launched by DARPA in 2017 [20].

Even though the general aim for XAI is well understood as the achievement of *interpretability*, or *explainability*, for ML models, few articulate precisely what those terms mean or why they are important. In this chapter, we try to provide a more rigorous definition for such terms, by reviewing what has been done in the literature so far.

2.1 Why do we need explainability

Before determining a good definition for *explainability* or *interpretability*, we must have a good understanding of what the real world objectives of XAI research are. More specifically, what are the desiderata of XAI which are still not being satisfied by the current ML tools and practices. Consider a supervised learning scenario: a lot of evaluation metrics are used to assess the quality

of a model, accuracy probably being the most common one. The computation of such metrics require the presence of model predictions, together with ground truth labels, in order to produce a score which is computed in order to answer in a quantitative way to some questions, like "how good is Model A able to generalize with respect to Model B?", or "what is the probability that Model A will misclassify an unseen sample?". This evaluation framework provides satisfactory answers for some kinds of questions, but still fails to answer other ones, especially those demanding qualitative information, with questions like "why did the model predict sample x to belong to class k?". This kind of questions are the ones sought by XAI research. However, we argue that a rigorous definition for the desiderata of XAI cannot consist in a list of potential questions: this approach is qualitative and already tackled by philosiphical works [21]. We therefore need to express such needs in other forms other than simple questions.

Lipton [22] suggests that the need for explainability arises "when our real world objectives are difficult to encode as simple real-valued functions": in this sense, *interpretations* are useful to achieve objectives which are important for us, but which we struggle to model in a formal way. Other mentioned motivating aspects are causality, transferability, informativeness and fair and ethical decision making. The authors in [23] refer to this same concept as *incompleteness* in the problem formalization. In such situations, explanations can act as a bridge between the model and the human supervisor, who can evaluate the predictions based on the provided explanations and decide whether they meet certain criteria that the machine alone could not understand. Some examples of such scenarios would be:

- **Scientific Understanding**: humans learn about the world around them in the form of knowledge, which is still difficult to formalize in the same way in which it works inside our brains. For this reason, we might look for explanations from ML models, which in turn we can interpret and transform into human interpretable knowledge.

- **Safety**: in complex tasks, rigorous and complete testing is almost never feasible and it is not possible to formally model all the wrong or dangerous decisions that a model could make. For this reason, when decisions from a ML system can pose a threat to others, explanations can help humans to evaluate safety conditions and to boost trust towards AI.

- **Ethics**: for us humans, evaluating the fairness of a decision is often easy, in the same way in which we have a clear idea of how we would want our model to be ethical (for example, we could desire a "fair" classifier for loan approval). However, this kind of properties are not easy to encode in ML systems and, at the same time, biases in the data can often lead to unethical decisions if not treated properly.

Some papers [24, 25] also motivate the need for explainability and interpretability in light of the need for trust by domain experts: indeed trust is fundamental if one plans to act based on a prediction, therefore ML systems must be able to communicate with highly skilled human experts to leverage their expertise and share useful information or patterns from the data.

2.2 When do we need explainability

Explainability is important in a lot of domains, but not in all of them. There are applications, e.g. aircraft collision avoidance, in which algorithms have been functioning from years without giving any explanations and without any human interaction. It is clear, then, that ML systems can be used without any need for interpretations in real world applications, at least in those cases where their raw performance in terms of accuracy suffices, or when the risk of error doesn't pose any serious threat to the end users. Therefore, domains that demand explainability are generally characterized by the critical nature of decisions which need to be made, where mistakes could have severe consequences. Authors in [26] provide a fairly exhaustive list of domains in need for explainability:

- **Medical Domain/Health-Care**: when the lives of humans are at stake, the need for explanations and knowledge are of paramount importance; take as an example a model used from doctors in order to associate to each patient a risk of suffering a certain disease. Such a model should not only be accurate, but intelligible: in this way, doctors could understand the underlying causes for such a disease, effectively introducing new scientific knowledge in the medical domain;

- **Judicial System**: machine learning systems have also been explored for the automatic decision of judgement results [27]. Such systems should help judges and lawyers to take decisions, but in order to do so their decision must be well motivated;

- **Banking/Finance**: typical examples of automatic decision making in the banking domain are automatic credit approval systems. Since banks are legally obliged to provide customers with motivations when their credit request is denied, the usage of explainable models is required;

- **Automobile Industry**: autonomous driving systems are one of the most common and popular applications in DL research. Such autonomous agents are responsible for any accident that could take place on the road: for this reason explanations for each of the agent's decision must be provided, both for legal and security reasons, so that the system can be quickly fixed and improved;

- **Recommender Systems**: explainable recommendations boost the trustworthiness and effectiveness of recommender systems [28]. Furthermore, regulations like the aforementioned GDPR are currently requiring models that use users' personal data to provide predictions, to also provide explanations.

Examples of other domains requiring explainability include bio-informatics, marketing, election campaigns, precision agriculture or expert systems for the military.

2.3 What is explainability

Several research works attempt to describe rigorously the meaning of explainability in the contest of XAI. In the literature, such word is often used interchangeably or substituted by "intrepretability", even though some try to make a distinction. In [29], for example, the authors claim that explainability is a property of a model that implies interpretability, but not viceversa. More specifically they provide definitions that, we argue, are the most agreed upon in the literature. For clarity, from now on we will interpret those two terms with the following meanings:

Definition 2.3.1 (Interpretability)**.** We define *interpretability* in the context of supervised ML as the generic property of a model which makes its individual components, as well as its functioning as a whole, understandable by a human being. Examples of interpretable models are simple linear regression or decision trees. Examples of not interpretable models are deep neural networks.

Definition 2.3.2 (Explainability)**.** We define *explainability* in the context of supervised ML as the generic process by which we extract any kind of explanation from a model. This can be done by exploiting the natural properties of the model (in which case, such a model can be called *explainable*), or by devising techniques to extract explanations from any model.

Explanations, according to the authors, can be evaluated in two ways: according to their intelligibility (that is, its understandability by a human being) and their completeness (that is, the accuracy of the description). Under this definitions, the challenge of XAI is in creating explanations that are both interpretable and complete, even though such characteristics are often opposed to one another. These two features of explanations resemble two important properties of ML models and suggest a similitude: on one hand, the user would desire a simple model with few parameters and at the same time a model able to capture really well the structure

of the training data. While both the properties are desirable, they are almost never achievable at the same time. Indeed, as we train simple models, we will probably underfit the data: in the same way, really easy explanations often fail to capture the complexities behind the internal workings of our algorithms. On the other hand, as we add parameters to our model and make it more complex, it will begin to better fit the data: in the same way a really complete explanation will describe accurately the operations of the systems, but will probably result more difficult to understand to a human being. This comparison suggests that, even for explanations, one should allow for a *tradeoff* between interpretability and completeness. The author also suggests that the explanation methods should be evaluated according to how such explanations behave on the curve from maximum interpretability to maximum completeness. This approach is followed in depth in [25], where the authors devise an explanation technique able to work for any classifier, by optimizing the intelligibility-completeness tradeoff. More information about this work is provided in Section 2.3.2.

In [23] interpretability is defined as the "ability to explain or to present in understandable terms to a human". According to [26], instead, *interpretability* is most often used in terms of comprehending how the prediction model works as a whole, while *explainability*, in contrast, is used to indicate the capability of models to give explanations about their decision, but keeping their black box nature. In the particular context of generating explanations for DL models, the authors of [30] define an explanation as the collection of features of an interpretable domain (i.e. pixel values on images), that have contributed for a given example to produce a decision (e.g. classification or regression). We can see that none of the aforementioned definitions are specific enough to enable one universal formalization: indeed they implicitly depend on the context or the aim of the research work.

Going back to our definitions, we could be tempted to consider an inherently explainable model generally more desirable with respect to an interpretable model. Indeed, interpretability implies an active effort on the part of the human supervisor to dig inside the model, understand the internal mechanisms and infer the model motivations. On the other hand, this process is simplified if the model provides explanations directly, allowing us to immediately have more insights about its decision process. While this may often be true, this certainly does not constitute a rule and, in general, a model can't be said to be interpretable if it is considered explainable, or vice-versa. For example, think about the human brain: while we are able to give really detailed and motivated reasons behind our decision processes, our brain is not an interpretable model. Indeed, we don't know every single aspect of how our brains works, and yet we (often) trust the explanations that other human beings provide when asked why they took a certain

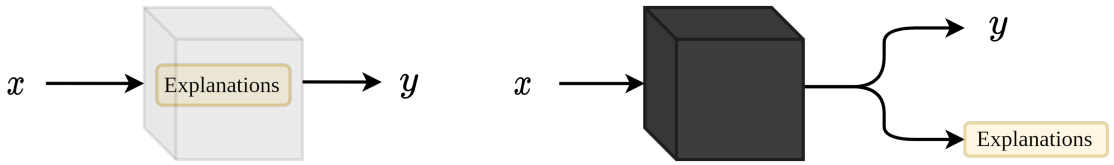


Figure 2.1: Intuitive representation of explanations coming from transparent models (left) vs. explanations coming from post-hoc interpretability techniques (right). Transparent models are inherently interpretable by their design; they can be inspected and explanations can be deduced from them. Post-hoc interpretability, on the other hand, refers to a wide range of techniques with which explanations are extracted by any already trained model.

decision. This offers an interesting point of view to the discussion: we should be careful not to trust certain explanations only by the fact that they look plausible and convincing. To this regard, Herman [31] warns its readers, making a clear distinction between descriptive and persuasive explanations: indeed implicit cognitive biases of the human brain could mislead us to trust wrong explanations (for example, humans naturally tend to prefer simpler descriptions). To avoid falling in this problem, one should always keep in mind the intelligibility-completeness tradeoff mentioned above.

It is clear that, even after providing some sort of definitions, the meanings of interpretability and explainability are still very generic and slippery. Nevertheless, the volume of research in XAI is quickly expanding, making the number of available methodologies continuously grow. One fundamental problem in XAI is the definition of specific properties, which make models explainable or interpretable. In the literature, two paradigms are often distinguished [32, 22]:

- **Transparency**, or integrated interpretability: it is mostly a feature of *interpretable* models. A transparent model can be interpreted by a human thanks to its own easy to understand design.
- **Post-hoc** interpretability: refers to that approach with which explanations are extracted from already trained models. Such models can be transparent, or retain their black box structure.

An intuitive illustration of those two concepts are illustrated in Figure 2.1.

2.3.1 Transparency

Transparency is one of the properties that can enable interpretability and it implies some sort of understanding of the mechanism by which the model works. It can also be seen as the direct opposite to the concept of *black box*. Lipton [22] goes into even more details, by subdividing transparency in different levels:

1. **Simulatability**: it's the highest level of abstraction of the concept of transparency. Lipton refers to simulatability as the property of the model that makes it understandable by a person "at once". Specifically, this means that a human could, given the input data and all the necessary parameters, produce a prediction by making all the computations in a reasonable time. This notion of transparency is not very applicable to modern machine learning techniques and is often disregarded.

2. **Decomposability**: it's the transparency considered at the level of the single components of the model. Specifically, one model can be considered decomposable if each part of the model (weights, modules, computations...) admits an intuitive explanation.

3. **Algorithmic Transparency**: this notion of transparency refers to the learning algorithm itself. For example, we know that in the case of linear regression the shape of the loss function is known, as well as an analytical form for the solution for the problem. This means a maximum degree of algorithmic transparency. On the other hand, modern deep learning lacks this notion of transparency: in fact, even if a lot of powerful optimization algorithms give empirically excellent results, there is no guarantee that those will work on any new problem. The same holds for the shape of the error function, which is almost never known.

It's interesting to notice that the human brain, as noted previously, doesn't exhibit any of those features. In fact, human thinking is not transparent to us and justifications in the form of explanations may differ from the actual decision mechanism. Transparent models are fascinating, but recent research in DL has proven that predictive performances rise when building deeper models, and not vice-versa. For this reason, in recent years, a lot of attention has been put on the research of post-hoc explainability methods.

2.3.2 Post-hoc interpretations

With post-hoc interpretability we refer to those methods with which we generate explanations from already trained models, without caring about their internal mechanisms. The advantage of this approach is that it does not impact on the performances of the model, which is treated as a black box. Unlike transparency, this kind of interpretability is the one that applies to humans. Lipton [22] summarizes post-hoc explainability methods into the following categories.

Visualization

Another popular approach for post-hoc interpretations is to provide visualizations in order to have a qualitative idea of what the model has learned. For example, when models learn em-

beddings in high dimensional vector spaces, one popular technique to visualize them is t-SNE [33], which provides 2D or 3D visualization of high dimensional data points, in such a way that nearby samples are likely to appear closer together. In the field of computer vision, several papers have investigated the internal representations of visual concepts in CNNs. In [2], the authors try to build "prototipes" of certain objects starting from already trained image classification models. Specifically, starting from a white noise image, they tweak it in such a way that the activation of a certain neuron (in this case, the output neuron corresponding to the selected object) is maximized. This process can be replicated also starting from an existing image, a process which produces really peculiar visual effects (see Figure 2.2).

Figure 2.2: Process of image manipulation shown in [2]. From an intuitive point of view, starting from existing real images, the network is asked to enhance the features of the image that resemble the desired object (in this case, starting from an image of a tree, start looking for features that resemble buildings). This process is then repeated, creating a positive feedback loop. As a result, the features of the desired objects appear seemingly out of nowhere.

This process is also known as *Activation Maximization*: consider a deep NN classifier which maps an input tensor (in this case an image) x to a set of classes $\{\omega_i\}_{i=1}^m$. In a classification scenario, we know that the i-th output neuron encodes the modeled class probability $p(\omega_i|x)$. The basic idea is that the *prototype* x_i^*, representative of class ω_i can be found as follows:

$$x_i^* = \max_x \log p(\omega_i|x) - \lambda\|x\|^2.$$

The proposed definition doesn't yield good results in practice: although producing strong class response, they often look unnatural. This problem is solved in several ways, for exam-

ple by adding regularization via a data density model, or imposing prior constraints which are common in real images, such as high correlation among neighboring pixels. Again, a similar approach is found in [34], where the authors investigate the amount of information retained in the hidden representations of CNNs. They manage to reconstruct the original images with good accuracy even from high level representations by performing gradient descent on white noise inputs.

Local explanations

With visualization techniques, one aims to understand what the model learned from a global point of view: for example, the activation maximization technique described earlier will reflect the "internal representation" that the model has of a particular class. With local explanations, instead, one aims at giving explanations in the context of a specific example. One popular approach is the computation of saliency maps, or relevance scores [35, 30, 36]. Specifically, given a data point $x \in \mathbb{R}^k$, and given the learned function f, the predicted class for point x will be $f(x)$; this approach aims to assign to each feature a score $R(x)_i$, representing a measure of how relevant the feature x_i is for explaining $f(x)$. One simple approach to obtain the relevance scores is *sensitivity analisys*: it consists on finding the input features for which the output is more sensitive, meaning the ones that best contribute to increase the value of the output. In mathematical terms, it is defined as:

$$R(x)_i = \left(\frac{\partial f}{\partial x_i} \right)^2,$$

where the gradient is then evaluated on x. Such relevance scores can then be visualized: for example, in images they can be represented as a mask. We should note a subtle detail: this method gives us an explanation of the *variation* of $f(x)$, not of the value of $f(x)$ itself. In other terms, this method does not answer the question *"what parts of x make it belong to class y?"*, but instead it answers *"what parts of x make it belong more/less to class y?"*.

While the relevance scores can be seen as a way of extracting explanations, those can also be used in different ways. An interesting approach can be found in [37], where the authors explicitly train the relevance scores in a supervised manner, in such a way to make them conform to some manually curated notion of where the attention should be put. Following this approach, the authors aim to train a model which is "right for the right reason". The motivation is that, if the assumption that relevance score faithfully describes the model's underlying behavior, then costraining such relevance scores in order to match domain knowledge would result in a model

that, in some way, will use that domain knowledge to take decisions.

Another interesting approach for generating local explanations is from Ribeiro et al. [25]: they propose an explanation technique called LIME (Local Interpretable Model-agnostic Explanations), which purpose is to locally approximate complex models with simpler, interpretable models, following the completeness-interpretability tradeoff mentioned by Gilpin in [29] . More specifically, they define an explanation as an interpretable model $g \in G$, where G is a set of models which, with the right conditions, can be considered interpretable (e.g. decision trees). In fact, the authors note that even those that are considered transparent models, will not be considered interpretable by a human supervisor if the model relies on thousands of features to make the prediction. This means that there should be a measure of complexity that determines how much a "potentially interpretable" model g is actually interpretable. They denote such measure of complexity as $\Omega(g)$ (e.g. for decision trees, this could represent the depth of the tree). Then, they denote with $f : \mathbb{R}^k \rightarrow \mathbb{R}$ the model to be approximated, and with $\pi_x(z)$ a proximity measure between a feature vector x and z. Finally, $\mathcal{L}(f, g, \pi_x)$ is defined as a measure of how badly the model g approximates f near the input sample x. With this setup, one can find a model g which is maximally interpretable, and which best explains the original model, near a specific input. Specifically, this model is the output of LIME, which is defined as follows:

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \mathcal{L}\left(f, g, \pi_x\right) + \Omega(g).$$

Another recent technique which provides local explanations is the attention mechanism [38], which, in recent years, has proven to be very successful mostly in machine translation [39] and computer vision [40]. Despite not being explicitly designed to provide explanations, they can do so by visualizing their internal attention scores, which highlight sections of the input data which were most influential for the classification. It is interesting to note that this kind of explanations are almost identical to the ones provided by saliency maps, discussed above. The core difference is that, for the attention mechanism, the attention scores are directly computed by the model at inference time, while the saliency maps are obtained after inference via backpropagation. It is interesting to know that datasets depicting how humans distribute attention have been created [41, 42]: this can allow us to evaluate attention-based models according to how much their attention patterns conform with the human ones. Notice how this particular framework differs a bit from standard post-hoc explainability methods; indeed, since the attention weights are jointly trained with the model, there is no need for any extra step after training

to obtain them, they just need to be extracted from the trained model. This might lead us to put them in the set of transparent models, but it should be noted that models based on attention are often very deep, are largely made up of standard NN layers, and the attention weights are trained jointly with all the other parameters of the model via backpropagation. Thus, they definitely lack both the simulatability and algorithmic transparency aspects. In the end, these models don't perfectly fall into any category, but since they only provide local explanations, we argue that this is the most appropriate way to describe them.

Explanation by example

A common way in which humans justify their decisions is by offering analogies between the current object in study, and similar objects. For example, to decide the best treatment for a medical condition, doctors often refer to previous similar case studies. This kind of explanation is referred to explanation by example. The first approach with this method was proposed in [43]: the authors aim to generate explanations in the form of a collection of elements from the training set, returned by the model (in this case, a neural network) together with the actual prediction. To do this, they generate a distance metric directly from the model: in this way it is possible, at inference time, to scan the entire training set and look for the data points which, by the model's point of view, are the most similar to the current one being classified. Given x the current sample being predicted and given y any element of the training set, this metric is defined as the euclidean distance between the hidden representations of y and x.

2.4 Neural Symbolic Integration

Despite the attempts in the literature to give rigorous definitions of explainability and interpretability, those remain slippery concepts: interpretations may vary depending on the application domain and type of task to be performed. The need for explainability is mostly due to the black box nature of ML models, which is universally agreed upon. Indeed, differently from humans, ML models employ only sub-symbolic representations of concepts, meaning that everything inside the model is encoded as tensors of real values, which are inherently opaque to humans. Furthermore, classic ML systems learn everything from data, which is a radically different paradigm from the one humans use for learning. This has been identified as one of their major flaws [4]: NNs can offer outstanding performances when the problem is confined inside a specific domain (e.g. recognize spoken words in a short audio clip), but have nothing to offer

when it comes to trivial tasks like commonsense reasoning. One possible way to extend the range of applications of NNs, could be to equip them with the ability to reason with abstract symbolic terms. Indeed, the introduction of symbolic language inside neural networks could also help in the process of encoding general knowledge inside them, a problem for which the formulation is still considered incomplete (cfr. Section 2.1). Furthermore, the integration of the statistical nature of learning with the logical nature of reasoning has been already identified as one of the most important research challenges in computer science [44].

One of the first proposed ML models able to handle uncertainty and logic simultaneously are Markov Logic Networks [45]. The authors, motivated by the growing interest in Statistical Relational Learning, propose a simple and effective representation able to combine the power of probabilistic graphical models with logic. Specifically, a MLN is defined as a set of pairs (F_i, w_i), where F_i is a First Order Logic (FOL) formula, and w_i is its corresponding weight. Those, together with a set of constants $C = \{c_1, \ldots, c_{|C|}\}$, are used as a template to build a Markov Random Field [46], which models the distribution over all the possible worlds as follows:

$$P(X = x) = \frac{1}{Z} \exp\left(\sum_i w_i n_i(x)\right),$$

where $n_i(x)$ is the number of true groundings of the i-th formula, in the current world. This simple and elegant hybrid approach makes it possible to handle the randomness and uncertainty of real-world problems together with the power of logic. However, MLNs can only handle discrete variables and features, which is a big limitation for real use cases. For this reason, an extension of MLNs has been proposed in [47], which can also deal with continuous variables.

However, with the recent success of DL models, a great interest has arisen in the integration of neural architectures with logic. The area of research that addresses the problem of integrating symbolical knowledge with neural architectures is known as Neural Symbolic Integration (NeSy). The intuition that motivates NeSy as field of research goes beyond explainability: while neural networks give good evidence to be a good modeling framework for the human mind (e.g. parallelization and adaptive learning from the environment) they completely lack the ability to reason in symbolic terms [48]. Indeed, works in NeSy argue that logic is the best tool to represent general knowledge, and their first aim is to build systems capable of dealing with both symbolic and sub-symbolic representations.

In the next chapter, we delve in the details of Knowledge Enhanced Neural Networks (KENN) [1], a neural network layer designed to integrate logical knowledge in NNs. We will also give a

short summary of other notable methods in NeSy and provide a comparison of experimental results on the task of collective classification.

3

Knowledge Enhanced Neural Networks

Knowledge Enhanced Neural Networks (KENN) [1] is a special type of NN layer, designed for injecting logical knowledge into a pre-existing base NN. More specifically, it is a residual layer designed to be stacked after the last layer of any NN, in order to boost its predictive performances via the addition of a Prior Knowledge in the form of first order logic clauses. In this chapter we will describe the theory behind KENN, its architecture, and experimental results.

3.1 Theoretical Framework

We present here the theoretical framework behind KENN. The first step will be to rigorously define the symbolic language and how to link it with the theoretical framework of NNs, which consists in defining a semantic for the language. Next, we will describe the process with which the truth value of a clause can be increased, and how to integrate this method inside NNs.

3.1.1 Prior Knowledge and language semantic

Definition 3.1.1 (Prior Knowledge)**.** The Prior Knowledge is defined as a collection of formulas of a function-free FO language $\mathcal{L}$ whose signature is defined with a set of constants $\mathcal{C} = \{a_1, \ldots, a_{|\mathcal{C}|}\}$ and a set of predicates $\mathcal{P} = \{p_1, \ldots, p_{|\mathcal{P}|}\}$. A predicate is a symbol, which can take in input a number of constants $n \geq 1$. If $n = 1$, the predicate represents a

property of the input constant; if $n > 1$, it represents a relation between the input constants. The number n is called the *arity* of the predicate.

Definition 3.1.2 (Clause)**.** A clause is defined to be of the following form:

$$c := \bigvee_{i=1}^{k} l_i, \quad l_i \neq l_j \quad \forall i \neq j. \tag{3.1}$$

In this equation, l_i is a literal, i.e. a formula constituted only by a n-ary predicate, or its negation. Specifically, when writing $l_i \neq l_j$, we mean that those two literals never share their constituting predicate. For example, the clause $A(x) \vee \neg B(x)$ is considered an acceptable clause, while $A(x) \vee B(x) \vee \neg A(x)$ is not. Also clauses have an arity, which is by definition the number of variables given in input to the clause.

One example of a clause could be the following:

$$c(x, y) = \neg Smoker(x) \vee \neg Friends(x, y) \vee Smoker(y), \tag{3.2}$$

which, in classical logic, is equivalent to the clause $Smoker(x) \wedge Friends(x, y) \Rightarrow Smoker(y)$, but expressed as a disjunction of literals. Such a clause is constituted by two predicates: $Smoker(x)$, a unary predicate expressing the statement "x *is a smoker*", and $Friends(x, y)$, a binary predicate which expresses the statement "x *and* y *are friends*". Therefore, this clause expresses the rule "*if* x *is a smoker and* x *and* y *are friends, than also* y *is a smoker*". Note that the variables x and y are supposed to be universally quantified, since our aim is to express general knowledge. We now give another definition:

Definition 3.1.3 (Grounding of a clause)**.** The grounding of an n-ary clause c, denoted as $c[x_1/k_1, \ldots, x_n/k_n]$, is the clause obtained by substituting k_i to $x_i, \forall i = 1, \ldots, n$.

Going back to the example of before, assume that a and b are two specific persons. Then, the grounding of clause (3.2) will be

$$c(x/a, y/b) = c(a, b) = \neg Smoker(a) \vee \neg Friends(a, b) \vee Smoker(b).$$

The next step is to build a semantic for the formal language $\mathcal{L}$, that is, how to interpret the symbols that we are working with. In practice, this will consist on defining a way to map constants towards a domain, and predicates to functions that go from such domain to a truth

value. To clarify, consider the following example: let a be a constant and let P be a predicate, such that $P(x)$ expresses the statement "x *is a prime number*". In this case, there is a natural way to define an interpretation for our symbols, that is to map constants to the domain of natural numbers and to map P to the function $f : \mathbb{N} \longrightarrow \{0, 1\}$, where $f(n) = 1$ if n is prime, and 0 otherwise. Now, we define the semantic of $\mathcal{L}$.

Definition 3.1.4 (Semantic of $\mathcal{L}$)**.** The semantic of $\mathcal{L}$ is defined by means of a pair of functions $(\mathcal{I}_\mathcal{C}, \mathcal{I}_\mathcal{P})$, that, together, define an *interpretation* for the symbols of our language and are defined as follows:

$$\begin{aligned} \mathcal{I}_\mathcal{C} : \mathcal{C} &\longrightarrow \mathbb{R}^l & \qquad \mathcal{I}_\mathcal{P} : \mathcal{P} &\longrightarrow \left(\mathbb{R}^{nl} \to [0, 1]\right) \\ c &\longmapsto x, & \qquad P &\longmapsto f \end{aligned} \tag{3.3}$$

Where n is the arity of the predicate P and f is a function that takes in input the concatenation of the interpretations of n constant symbols, $\mathcal{I}_C(c_1), \ldots, \mathcal{I}_C(c_n)$ and returns the truth value of $P(c_1, \ldots, c_n)$. Note that, to make the notation lighter, we will omit the subscript when it's clear whether the argument of the interpretation is a literal or a constant term.

With those definitions, we can already see how this theoretical framework is suitable to describe a NN. In fact, each constant symbol c is mapped to a l-dimensional real vector, which can be seen as the feature vector characterizing the real world object identified by c. Another important detail is that the truth value of each literal, in our setup, is not determined by a hard assignment of 0 or 1, but is represented by a real number in the interval $[0, 1]$. This is a crucial point: indeed, the truth value in our semantic represents predictions of a NN, which, for classification tasks, are always expressed in terms of values in the interval $[0, 1]$. The natural consequence of this choice is that, from this point on, we will have to rely on the rules of Fuzzy Logic [49], which is a generalization of the standard Boolean logic where the truth value of variables can take the value of any real number between 0 and 1.

3.1.2 t-conorm Functions

With our definition of a semantic for $\mathcal{L}$, we can now give an interpretation for constants and predicates. The next step is to find a way to interpret clauses, or, more specifically, a way to determine the truth value of a grounded clause. We saw that, by definition, a clause is a disjunction of literals: this means that we only need a way to define the interpretation of a negated predicate and of the disjunction of two predicates. As stated above, since we are allowing truth

values in the range $[0, 1]$, we will need to use the rules of Fuzzy Logic. For computing the truth value of a negated predicate, the standard way in Fuzzy Logic is to use the Lukasiewicz Negation.

Definition 3.1.5 (Lukasiewicz Negation)**.** If $P \in \mathcal{P}$ is a predicate, then:

$$\mathcal{I}(\neg P) = 1 - \mathcal{I}(P)^* \tag{3.4}$$

So for example if the truth value of a predicate is $\mathcal{I}(P)(x) = 0.8$, the truth value of its negated copy would be $\mathcal{I}(\neg P)(x) = 0.2$. It is worth noting that this definition is equivalent to the Boolean negation when $\mathcal{I}(P) = 0$ or $\mathcal{I}(P) = 1$.

With this tool we are now able to compute the truth value of any literal. There remains to see how to define the interpretation of a disjunction of literals. To do this, we introduce the concept of t-conorm functions.

Definition 3.1.6 (t-conorm)**.** A t-conorm is a function $\perp\colon [0, 1]^2 \to [0, 1]$ that satisfies the following properties:

1. $\perp (a, b) = \perp (b, a)$
2. $\perp (a, b) \leq \perp (c, d)$ if $a \leq c$ and $b \leq d$
3. $\perp (a, \perp (b, c)) = \perp (\perp (a, b), c)$
4. $\perp (a, 0) = a$

By definition, $\perp$ takes values in $[0, 1]^2$, but can be easily extended to $[0, 1]^n$ for any n, by defining:

$$\perp (a_1, \ldots, a_n) := \perp (a_1, \perp (a_2, \cdots \perp (a_{n-1}, a_n))).$$

In Fuzzy Logic, t-conorm functions are used to represent the concept of logical disjunction, and will be the tool employed to represent the interpretation of a disjunction of literals. Specifically:

$$\mathcal{I}(l_1 \vee \cdots \vee l_n) = \perp (\mathcal{I}(l_1), \ldots, \mathcal{I}(l_n)). \tag{3.5}$$

*Writing $1 - \mathcal{I}(P)$ is a slight abuse of notation since $\mathcal{I}(P)$ is a function. To be more precise one shoud write $\mathcal{I}(\neg P)(\mathcal{I}(c)) = 1 - \mathcal{I}(P)(\mathcal{I}(c)), \forall c$.

With the given definitions, we have all that is needed to compute the truth value of any grounded clause. From a practical point of view, the only remaining step would be to choose a specific t-conorm function. KENN uses the Gödel t-conorm function, which is also known as the Maximum t-conorm and is defined as

$$\perp_{max}(a, b) = \max\{a, b\},$$

which, as above, can be extended like follows:

$$\perp_{max}(t) = \max_{i=1,\dots,l} t_i, \quad \forall t \in \mathbb{R}^l.$$

We are now finally ready to fully understand how this theoretical framework is able to describe the predictions of a NN. Suppose that we have a dataset $\mathcal{X} = \{x_1, \dots, x_n\}, x_i \in \mathbb{R}^l$, where each x_i belongs to one or more classes $(P_1, \dots, P_m)$. The task in which the NN must learn to classify each input into one or more output classes is known in Machine Learning as a multilabel classification problem. To tackle this kind of task, a NN architecture will present, in the last layer, m output units, each of which will be finally subject to a sigmoidal activation function. After training, the NN will have learned to approximate a function $h(x_i) = y_i \in \mathbb{R}^m$, where $(y_i)_j = \mathbb{P}(x_i \text{ belongs to class } j)$. Now, if we consider:

1. $\mathcal{P} = \{P_1, \dots, P_m\}$ to be predicates defined as $P_i(x) =$ "x belongs to class P_i";
2. $\{x_1, \dots, x_n\}$ to be the interpretations of the constants $\mathcal{C} = \{c_1, \dots, c_n\}$, which represent the real-world objects of our dataset,

it is clear that the entries of y_i can be seen as truth values of the predicates $\{P_1, \dots, P_m\}$. More formally:

$$(y_i)_j = \mathcal{I}_{NN}(P_j)(x_i), \quad \forall i = 1, \dots, n, \forall j = 1, \dots, m. \tag{3.6}$$

Hence, the whole NN defines an interpretation for each predicate P_i, which we denoted as $\mathcal{I}_{NN}$. Therefore, given a clause $c := \bigvee_{i=1}^{k} l_i$ and given a collection of feature vectors $\{x_1, \dots, x_d\}$ (where d is the arity of c), then the truth value of the grounded clause predicted by the NN will be $\perp((y_c)_1, \dots, (y_c)_k)$, where:

$$y_c \in \mathbb{R}^k, \quad (y_c)_i = \begin{cases} \mathcal{I}_{NN}(l_i) \text{ if } l_i \text{ is not a negated predicate} \\ 1 - \mathcal{I}_{NN}(l_i) \text{ otherwise.} \end{cases} \tag{3.7}$$

The intuition behind KENN is very simple: given the vector y of predictions of the NN, a new layer is added at its end with the aim to modify y and obtain a new vector of predictions y', of the form $y' = y + \delta$, such that y' improves the truth value of each clause present in the base knowledge and, at the same time, keeps the quantity $\|y' - y\|_2$ minimal. It is worth noticing that this new layer introduced by KENN, called Knowledge Enhancer (KE), is a kind of residual layer, since it learns to represent the quantity $\delta = y' - y$.

3.1.3 t-conorm Boost Functions

The next problem is to understand how to improve the truth value of a single clause. Since this truth value is represented by a t-conorm function, this involves finding a way to let the value of $\perp (y)$ rise by manipulating the value of y. To do this, we define a new class of functions.

Definition 3.1.7 (t-conorm Boost Function (TBF))**.** A function $\delta : [0,1]^n \to [0,1]^n$ is a t-conorm Boost Function (TBF) if:

$$0 \leq t_i + \delta(t)_i \leq 1 \quad \forall n \in \mathbb{N} \quad \forall t \in [0,1]^n.$$

Let Δ denote the set of all TBFs.

From the definition follows a simple but essential result.

Lemma 3.1.1. Given any t-conorm $\perp$ and $\delta \in \Delta$, it holds that:

$$\perp (t) \leq \perp (t + \delta(t)).$$

Proof. By definition of TBF, $\delta(t)_i \geq 0$. This implies that

$$t_i \leq t_i + \delta(t)_i, \quad \forall i = 1, \ldots, n.$$

By the monotonicity of t-conorms (i.e. property 2 of Definition 3.1.6) , it follows that $\perp (t) \leq \perp (t + \delta(t))$. □

The purpose of such TBF δ is to update the NN predictions $y \in \mathbb{R}^m$ to a new vector in such a way that the truth value of the clause increases. The problem is now how to choose such a TBF. It is clear that not all the $\delta \in \Delta$ would be useful: for example, one could choose the function $\delta(y)_i = 1 - y_i, \quad \forall i = 1, \ldots, n$. In this way, we would obtain an updated truth

value of 1 for any literal, independently of y. This of course would be pointless, and would render the predictions of the base NN useless. For this reason another requirement for y' is needed. Specifically, as we already mentioned, KENN is built in such a way that the learnt δ improves the t-conorm value in a minimal way. To be more rigorous, we will now formally define the concept of a minimal TBF.

Definition 3.1.8 (Minimal TBF)**.** A function $\delta \in \Delta$ is minimal with respect to a norm $\|\cdot\|$ and a t-conorm $\perp$ if and only if:

$$\|\delta'(t)\| < \|\delta(t)\| \Rightarrow \perp (t + \delta'(t)) < \perp (t + \delta(t)), \quad \forall \delta' \in \Delta, \quad \forall n \in \mathbb{N}, \quad \forall t \in [0,1]^n.$$

As mentioned above, KENN works with the Gödel t-conorm function and the L_p norm $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{\frac{1}{p}}$. The next step at this point is to find such a minimal TBF. In the following result, we present a possible form that a minimal TBF can assume.

Theorem 3.1.2. For any function $f : \mathbb{R}^n \to \mathbb{R}$ we define $\delta^f : \mathbb{R}^n \to \mathbb{R}^n$ as

$$\delta^f(t)_i = \begin{cases} f(t) & \text{if } i = \arg\max_{j=1}^n t_j \\ 0 & \text{otherwise} \end{cases}$$

Let $f : [0,1]^n \to [0,1]$ satisfying $0 \leq f(t) \leq 1 - \max_{j=1}^n t_j$. Then, δ^f is a minimal TBF for the Gödel t-conorm function and the L_p norm.

Proof. δ^f is a TBF. Indeed $\delta^f(t) \geq 0$ and $0 \leq t_i + \delta^f(t_i) \leq 1$ because $f(t) \leq 1 - \max_j t_j$. Therefore we only need to prove that δ^f is minimal. Take $\delta \in \Delta$, with $\|\delta(t)\|_p < \|\delta^f(t)\|_p$. We have to show that

$$\perp (t + \delta(t)) < \perp \left(t + \delta^f(t)\right).$$

Now define $j = \arg\max_k (t_k + \delta(t)_k)$. By definition of the Gödel t-conorm we can immediately derive that:

$$\perp (t + \delta(t)) = t_j + \delta(t)_j. \tag{3.8}$$

Now, defining $i = \arg\max_k t_k$, using the same reasoning and exploiting the definition of δ^f it follows that:

$$\perp (t + \delta^f(t)) = t_i + f(t). \tag{3.9}$$

By combining (3.8) and (3.9) and noting that by definition $t_i \geq t_j$, the last step is to prove that $f(t) > \delta(t)_j$. To do this we exploit the definition of L_p norm as follows:

$$\delta(t)_j = (|\delta(t)_j|^p)^{\frac{1}{p}} \leq \left(\sum_{k=1}^{n} |\delta(t)_k|^p \right)^{\frac{1}{p}} = \|\delta(t)\|_p < \|\delta^f(t)\|_p = f(t).$$

Where the last inequality follows from the definition of $\delta^f(t)$.

□

3.1.4 Applying TBFs to preactivations

There is a problem with the definition of δ^f: there is a specific constraint $f(t) \leq 1 - \max_i t_i$ that limits the number of candidates for f. Indeed, this is imposed to ensure that the final output $y' = y + \delta^f(y)$ will be in $[0, 1]$. There is a natural way to solve this impracticality: since we are assuming a multilabel classification scenario, the final m output units of the NN will pass through a sigmoidal activation function. More specifically, y_i will be of the form:

$$y_i = \sigma(z_i) = \frac{1}{1 + e^{-z_i}}, \forall i = 1, \ldots, m.$$

For this reason, KENN exploits the fact that σ takes only values in $[0, 1]$ by applying the TBF directly on the preactivations $z \in \mathbb{R}^m$. In fact, it is clear from an intuitive point of view that one can apply any delta to the preactivations vector, and at the same time always be sure that the final output y will be in $[0, 1]$. In this way, the constraint on f is no longer needed. The next theorem proves formally that applying a minimal TBF on the preactivations z is equivalent to applying a minimal TBF on the output of the NN y.

Theorem 3.1.3. For all $f : \mathbb{R}^n \to \mathbb{R}$, the function:

$$\delta^g(y) = \sigma(z + \delta^f(z)) - \sigma(z) \tag{3.10}$$

is a minimal TBF under the Gödel t-conorm and the L_p norm.

Proof. By definition we know that $z = \sigma^{-1}(y)$, hence we can rewrite equation (3.10) as:

$$y + \delta^g(y) = \sigma(z + \delta^f(z)).$$

From the definition of sigmoid activation function it easily follows that $0 \leq y + \delta^g(y) \leq 1$. To prove that δ^g is a TBF we have to show that $\delta^g(y) \geq 0$. We note a few details: σ is monotonic

increasing, which means that:

$$\underset{j=1}{\operatorname{argmax}}^{n}_{j=1}\sigma\left(z_j\right) = \underset{j=1}{\operatorname{argmax}} z_j.$$

Another implication of the monotonicity of σ is that:

$$f(z) \geq 0 \Rightarrow \sigma\left(z_i + f(z)\right) \geq \sigma\left(z_i\right)$$

This implies that $\delta^g(y) = \sigma(z + \delta^f(z)) - \sigma(z) \geq 0$. We now define the function $g(y) = \sigma(z_i + f(z)) - \sigma(z_i)$, where $i = \arg\max_j z_j$. It's easy to see that this g is the function associated to our δ^g. In fact:

$$\begin{aligned}\delta^g(y)_i &= \sigma(z + \delta^f(z))_i - \sigma(z)_i \\ &= \begin{cases} g(y) & \text{if } i = \arg\max_j y_j \\ 0 & \text{otherwise.} \end{cases}\end{aligned}$$

Therefore, Theorem 3.1.2 guarantees that $\delta^g(y)$ is a minimal TBF under the Gödel t-conorm and the L_p norm.

□

Note that $\delta^g(y)$ is not directly used in KENN but it's indirectly induced by using $\delta^f(z)$ on the preactivations.

Applying the TBF directly on the preactivations has also another remarkable advantage. Indeed, it is known that it is possible to interpret the value of the preactivation of the i-th output neuron as the "confidence" of the NN that the current feature vector is to be classified in the i-th class. This "confidence" is not yet a probability, but a generic scalar value $z \in \mathbb{R}$; it will become a probability when transformed with the sigmoid activation function: $\sigma(z) \in [0, 1]$. More specifically we know that:

- $z \gg 0$ means high confidence of being classified in the i-th class. This follows from the fact that $\lim_{z\to+\infty} \sigma(z) = 1$;
- $z \ll 0$ means high confidence of *not* being classified in the i-th class. This follows from the fact that $\lim_{z\to-\infty} \sigma(z) = 0$;
- $z \approx 0$ corresponds to a highly uncertain decision. This follows from the fact that $\sigma(z) \approx 0.5$ if $z \approx 0$.

By observing the shape of the sigmoid activation function we can notice that when $|z| \gg 0$ (high confidence in the NN predictions), even large deltas on the preactivations produce very small changes. More rigorously, $\lim_{|z|\to\infty} |\sigma(z + \delta^f(z)) - \sigma(z)| = 0$. On the contrary, when $z \approx 0$, even small deltas on the preactivations produce high modification at the activation level. This will result in the following behavior: if the NN is highly confident of its decision, then logical rules will not modify too much the result of the NN predictions. On the contrary, in the cases where the NN is uncertain of its decision, our base knowledge will intervene and give higher modifications on the final predictions. This conforms to the intuition that KENN should produce minimal changes in the original predictions. These key concepts are further illustrated in Figure 3.1.

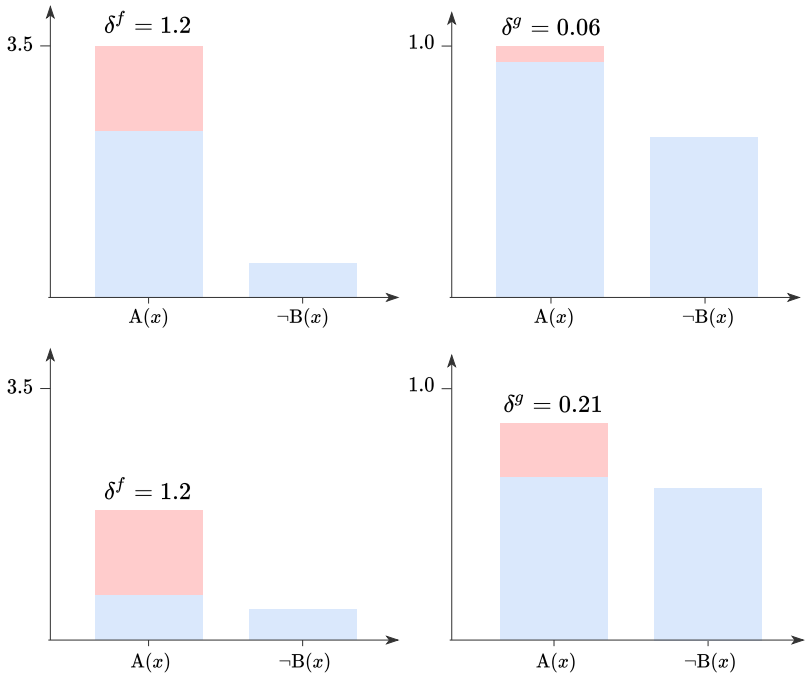


Figure 3.1: On the left, example preactivations for the clause $\mathrm{A}(x) \vee \neg\, \mathrm{B}(x)$ are shown. For both the examples, the same delta (δ^f) is applied to these preactivations. In the first example, the NN has a high confidence, while in the second one it is much lower. We can see how, when applying the activation function, the actual delta (δ^g) is much smaller in the first case and larger in the second. This is due to the shape of the sigmoid activation function itself.

As we already mentioned, the minimal TBF directly modeled by KENN is the one we called $\delta^f(z)$. From its definition, we know that the magnitude of the produced delta is determined by the definition of f. One of the most important features of KENN is that, by design, it learns to give the proper *importance* to each clause in the base knowledge: this precise feature

of the model gives also a way to find such function f. Specifically, for each clause c a learnable parameter w_c is defined so that the produced delta for c is:

$$\delta^{w_c}(z)_i = \begin{cases} w_c & \text{if } i = \arg\max_{j=1}^{n} z_j \\ 0 & \text{otherwise.} \end{cases}$$

From this definition it's now clear that the function f we were looking for is not actually the same for all the clauses in the base knowledge, but it is defined for each different clause and it's equal to the constant function $f_c(z) = w_c, \quad w_c \in [0, \infty]$. There is one last problem: while it's true that the function δ^{w_c} is a minimal TBF, the implementation of this kind of functions inside a NN is unfeasible since they are not differentiable. For this reason KENN uses a soft approximation of δ^{w_c}, defined as:

$$\delta_s^{w_c}(z)_i = w_c \cdot \text{softmax}(z)_i = w_c \cdot \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}. \tag{3.11}$$

There are still, however, some steps to describe in order to fully understand how KENN produces a vector of deltas. Recall our notation: we defined with $y \in \mathbb{R}^m$ the vector of predictions from the NN. Specifically, we now define with y_A the truth value relative to A, where A is a generic grounded predicate of our language. We also define $z_A = \sigma^{-1}(y_A)$. Now, we note that equation (3.11), tells us that the produced δ is always m-dimensional, where m is the number of output classes. This however is not desirable: in fact, in general, not all the clauses in our base knowledge contain all the predicates of our language. For example, given $\mathcal{P} = \{A, B, C\}$, the clause $c = A \vee \neg B$ contains only two of the three predicates in the language. Therefore, we would like this specific clause to not modify in any way z_C. Another problem is that we don't know how to express the preactivation of a negated literal, i.e. we don't know how to derive $z_{\neg A}$ from z_A. In fact, recall that in equation (3.4) we defined the interpretation of a negated predicate, where we knew that the truth values were well defined in the interval $[0, 1]$. Now we are dealing with preactivations, which cannot be considered truth values in the Fuzzy Logic theoretical framework. However, this problem can be easily solved by exploiting the following property of the sigmoid activation function:

$$1 - \sigma(x) = \sigma(-x).$$

Now it's easy to see that, since $y_{\neg A} = 1 - y_A$, we can define:

$$z_{\neg A} = -z_A.$$

Notice that we are not introducing any new concepts: instead we are just redefining quantities that were already mentioned at the activation level, to the preactivation level. We finally define $z_c = (z_{l_1}, \ldots, z_{l_k})$ for every clause $c := \bigvee_{i=1}^{k} l_i$ of the knowledge. We refer to the process of transforming from z to z_c as the *selection* step. This new vector contains only the preactivations of literals present in c, and is the one that we actually want to use to produce the delta relative to clause c. Now, let $\mathcal{K}$ be the set of clauses in our knowledge, and $\{w_c\}_{c \in \mathcal{K}}$ their corresponding weights. For every clause $c \in \mathcal{K}$ we want to obtain a new delta, namely $\delta^c \in \mathbb{R}^m$, which contains one value for each predicate in the clause and is defined as follows:

$$\delta^c_A = \begin{cases} \delta_s^{w_c}(z_c)_A & \text{if } A \in c \\ -\delta_s^{w_c}(z_c)_{\neg A} & \text{if } \neg A \in c\ , \quad \forall A \in c \\ 0 & \text{otherwise} \end{cases} \tag{3.12}$$

This newly defined delta, δ^c, will be the change obtained from clause c and will be summed to z to obtain the updated prediction. More specifically:

$$y' = \sigma(z + \delta^c)$$

In Figure 3.2, an example of computation of the δ^c vector is provided.

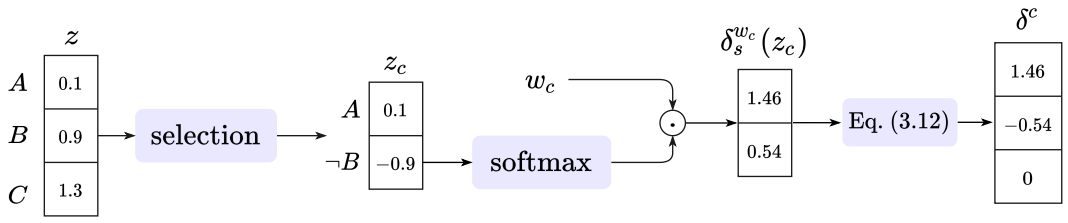


Figure 3.2: Example with all the steps needed to compute δ^c for the clause $A \vee \neg B$ starting from the vector of preactivations z; for this example $w_c = 2$. We refer to this process as *clause enhancement*.

3.1.5 Increasing the satisfaction of the Knowledge

In the previous section we found out how KENN produces a vector of changes δ to be applied to the original NN predictions, but only considering a single clause. That would suffice in the cases where the knowledge is constituted only by a single clause, but in real applications a higher number of logical rules will be desirable. Therefore, the next and final problem is to understand how KENN takes all the deltas from all the clauses and produces a single vector of changes. This particular step of aggregation is critical, as it constitutes one of the best features of KENN, but at the same time one of its bigger inaccuracies. This is because, to aggregate the contributions from all the clauses $c \in \mathcal{K}$, KENN just sums the contributions. Specifically, the final prediction is defined as follows:

$$y' = \sigma(z + \sum_{c \in \mathcal{K}} \delta^c). \tag{3.13}$$

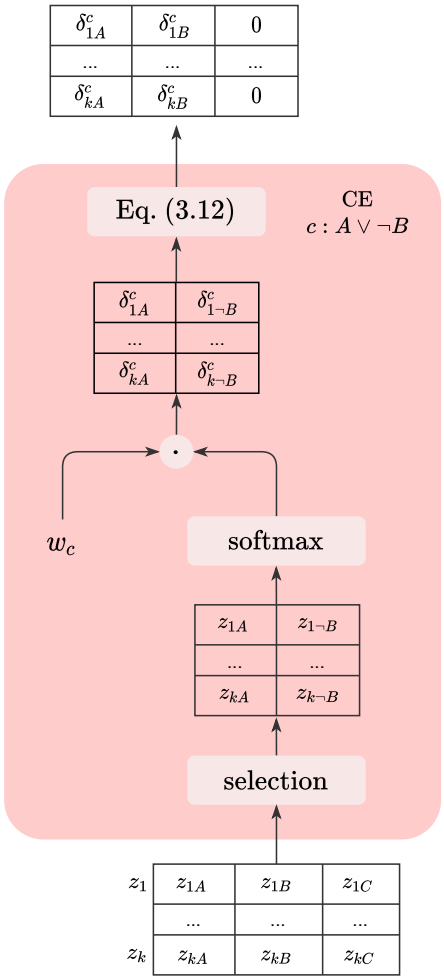
This particular choice makes KENN really fast at inference and learning time, increasing scalability. At the same time, though, this makes the risk of inconsistencies higher. For example, the same predicate can appear negated in one clause, and not negated in another clause: in this way the delta for the first one will be negative while it will be positive for the second one. In this way, the two deltas may cancel out rendering the contributions of the two clauses less effective.

3.2 KENN Architecture

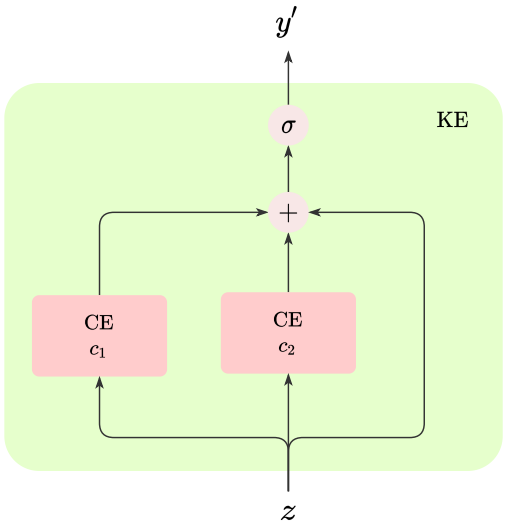
In this section we describe the architecture of the KENN layer in details. First, we will describe the basic functioning for the case where the knowledge is constituted only by unary clauses. Then, in Section 3.3, we will describe how KENN is able to handle binary clauses. Note that, theoretically, KENN can be extended to work with clauses of any arity. However, for now it has been implemented to work with unary and binary clauses, which are also the most common use cases. All the contents of this section have been implemented as a Python 3 package, based on TensorFlow 2 [50] and all the source code is publicly available on Github[†].

As described above, the core functionality of KENN is the *clause enhancement*, i.e. the creation of a vector δ which, summed to the vector of preactivations z from a NN, produces a modified vector of predictions y' which increases the truth value of the relative clause. The architecture that takes care of this task is the Clause Enhancer (CE): this submodule takes in

[†] `https://github.com/DanieleAlessandro/KENN2`



(a) The Clause Enhancer (CE) module.



(b) The Knowledge Enhancer (KE) module.

Figure 3.3: Illustration of the KENN architecture. Images are replications of the illustrations provided in [1].

input the full vector of preactivations from the original NN and computes the vector of deltas described in equation (3.12). For each clause in the base knowledge, a CE will be instantiated. Note how the CE does not take in input a single vector of preactivations at a time, but works in parallel taking in input mini-batches of data. The details of the architecture of the CE are described in Figure 3.3a. Specifically, the CE takes in input the mini-batch of preactivations $\{z_1, \ldots, z_k\}$ coming from the base NN, and for each sample in the mini-batch performs the following operations:

1. Performs the *selection* step;
2. Computes the vector of deltas according to equation (3.11);

3. Transforms the delta vector according to equation (3.12)

Once every CE has produced a vector of deltas, the next step is to aggregate them by summing them up. The module that performs this operation is called the *Knowledge Enhancer* (KE), which is shown in Figure 3.3b. More specifically, the KE performs the following operations:

1. Takes all the preactivations from the NN as inputs, and instantiates a CE for each clause in the knowledge base;
2. Sums the produced vector of changes from all the CEs, obtaining a final delta vector;
3. It sums the obtained delta with the original preactivations vector;
4. Applies the sigmoid activation function.

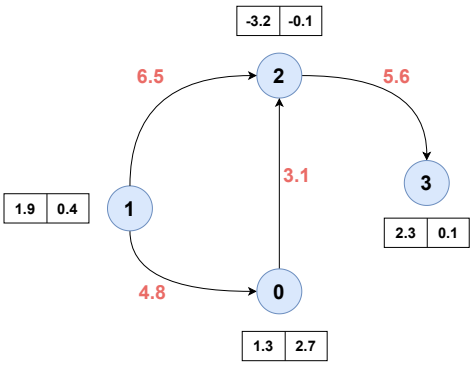
These steps produce the final vector of modified predictions y'. Notice that the KE is the implementation of equation (3.13).

3.3 KENN for relational data

As we said, KENN can also handle relational data; specifically, thanks to its general approach, KENN can theoretically handle clauses with any arity. Nevertheless, for simplicity, we will just describe how it is possible to handle the case of binary clauses. In order to integrate binary clauses, the first step is to split the base knowledge into two disjoint subsets of clauses $\mathcal{K}_U$ and $\mathcal{K}_B$, containing the unary and binary clauses respectively. In this way, the set of all the clauses constituting the Prior Knowledge is denoted as $\mathcal{K} = \mathcal{K}_U \cup \mathcal{K}_B$. The intuition at the base of KENN remains the same: starting from initial predictions z, we will need to produce a vector of deltas δ. The only difference, is that δ will be the sum of two different deltas, one produced just by unary clauses and the other just by binary clauses. More in details, this consists in modifying equation (3.13) as follows:

$$y' = \sigma(z + \sum_{c \in \mathcal{K}_U} \delta^c + \sum_{c \in \mathcal{K}_B} \delta^c). \tag{3.14}$$

Note that the procedure to compute the quantity $\sum_{c \in \mathcal{K}_U} \delta^c$ is the one we described before, since it deals just with unary clauses. The only thing left to define is how to compute a delta vector starting from a binary clause.



U

i	Unary	
0	1.3	2.7
1	1.9	0.4
2	-3.2	-0.1
3	2.3	0.1

B

sx	sy	Binary
0	2	3.1
1	0	4.8
1	2	6.5
2	3	5.6

Figure 3.4: Representation of relational data inside KENN. Specifically, objects and relations can be seen as nodes and edges of a directed graph. The preactivations of each grounded predicate are represented in tables U and B. In this example, two unary predicates and one binary predicate are present. Note that in matrix B only the object pairs such that there is a relation between them are reported.

Table representation of binary data

In order to deal with binary clauses, KENN represents truth values of every grounded predicate inside two different matrices: U and B. Matrix U has one row for each constant, one column for each unary predicate and an additional column containing an identifier for each object. Matrix B, instead, has one row for each couple of objects: specifically, for each pair of objects such that there is a binary relation between them. The first two columns of B contain the identifier indices of the pair of objects, while the remaining columns contain the preactivations for each binary predicate, grounded on that pair of objects. So, in the case of n unary predicates, m binary predicates, p total objects and q couples of objects, U and B will have dimensions $p \times (n+1)$ and $q \times (m+2)$ respectively. Figure 3.4 shows an example with a visualization of U and B. With those newly defined matrices, equation (3.14) can be implemented as follows:

$$\begin{cases} y'_u = \sigma(U + \delta U) \\ y'_b = \sigma(B + \delta B), \end{cases}$$

where y'_u and y'_b are the final predictions for the unary and binary predicates respectively, and δU and δB are the delta matrices that need to be computed.

The JOIN operation

At this stage, we have two matrices, U and B, containing preactivations of unary and binary predicates respectively. However, recall that the CE module takes in input a single matrix containing all the preactivations, in order to compute the deltas. The fact is that, given any binary clause, we must be able to access the preactivation value of any of its atoms, from a single matrix. Take as an example the grounded clause $A(c_1) \vee B(c_1, c_2) \vee \neg A(c_2)$. The preactivations of $A(c_1)$ and $A(c_2)$ can be accessed from matrix U, but the one for $B(c_1, c_2)$ can be accessed only from B. For this reason, we need a way to merge both the binary and unary preactivations into a single matrix. KENN does this with the JOIN operation.

Definition 3.3.1 (JOIN)**.** Given n unary predicates, m binary predicates and their respective matrices U and B, the JOIN operation is defined as follows:

$$\text{JOIN}(U, B) = M$$

where the i-th row of the matrix M is:

$$M_i = (s^x, s^y, U_{s^x 2}, \ldots, U_{s^x (n+1)}, U_{s^y 2}, \ldots, U_{s^y (n+1)}, B_{i3}, \ldots, B_{i(m+2)}),$$
$$\text{where } s^x = B_{i1}, s^y = B_{i2}.$$

More simply, M has the same number of rows as matrix B and shares its two first columns; additionally, it also contains the preactivations of all the unary predicates for both the first and second index of each row, together with all the preactivations for the binary predicates.

From this new matrix, the CE can access the preactivations for any atom of any grounded binary clause and can compute a delta matrix, which we will call δM.

The GROUP BY and SELECT operations

As we saw with our previous example, binary clauses can also contain unary predicates. This implies that the delta vectors generated from binary clauses can actually modify the preactivations of both unary and binary predicates. More specifically, this means that the matrix δM contains some information that is supposed to go both inside matrices δU and δB. KENN extracts the unary deltas from δM with the GROUP BY operation, and does the same for the binary deltas with the SELECT operation.

Definition 3.3.2 (GROUP BY)**.** The GROUP BY (GB) operation takes in input the matrix of deltas δM and an index (x or y), and is defined as follows:

$$\begin{cases} \mathrm{GB}(\delta M, x) = \delta U_x \\ \mathrm{GB}(\delta M, y) = \delta U_y, \end{cases}$$

where

$$\begin{cases} (\delta U_x)_i = \left(i, \sum_{j \in J_x^i} \delta M_{j3}, \ldots, \sum_{j \in J_x^i} \delta M_{j(2+n)}\right) \\ (\delta U_y)_i = \left(i, \sum_{j \in J_y^i} \delta M_{j(n+3)}, \ldots, \sum_{j \in J_y^i} \delta M_{j(2n+2)}\right) \\ J_x^i := \{j : \delta M_{j1} = i\} \\ J_y^i := \{j : \delta M_{j2} = i\}. \end{cases}$$

In simple terms, the GROUP BY operation extracts the deltas for the unary predicates from the delta matrix δM, by aggregating the results for each object identifier. This is exactly equivalent to a GROUP BY query in a relational database.

Definition 3.3.3 (SELECT)**.** The SELECT operation simply extracts the deltas for the preactivations of the binary atoms from the δM matrix. Specifically:

$$\mathrm{SELECT}(\delta M) = \delta B$$

where:

$$\delta B_i = \left(\delta M_{i1}, \delta M_{i2}, \delta M_{i(2n+3)}, \ldots, \delta M_{i(2n+m+2)}\right).$$

Computing the delta matrices

Now that all the necessary tools have been defined, we are ready to describe the process to compute the matrices δU and δB. KENN uses two different KEs to generate delta matrices: one (KE_u) takes in input the matrix U, and returns a delta matrix δU_u, using the standard procedure for unary predicates described in the previous section. The other (KE_b) takes in input a matrix $M = \mathrm{JOIN}(U, B)$, and returns a delta matrix δM. Afterwards, the following operations are performed:

1. The final delta matrix for binary predicates is obtained as follows: $\delta B = \mathrm{SELECT}(\delta M)$;
2. Two matrices $\delta U_x = \mathrm{GB}(\delta M, x)$ and $\delta U_y = \mathrm{GB}(\delta M, y)$ are computed;

3. δU_x and δU_y are summed together, obtaining matrix δU_b;

4. The final delta matrix for unary predicates is obtained as follows: $\delta U = \delta U_u + \delta U_b$.

In Figure 3.5, we report an example with the whole process.

3.4 Related Work

As already introduced at the end of Chapter 2, the integration of statistical learning with logical symbolic knowledge is one of the key challenges in the research on artificial intelligence, specifically for the field of NeSy. Authors of [1] subdivide NeSy sytstems into three main groups, based on their different objectives:

1. **Differentiable Reasoning** : in this category, the goal is to create differentiable approaches for deductive resoning, which can be defined as the process of producing logical deductions from an initial knowledge base;

2. **Inductive Logic Programming** : here the goal is to extract logical knowledge from data, or to refine an existing one;

3. **Knowledge Guided Learning**: here the goal is learning in classical ML sense, where the knowledge has the role of a supervisor, meaning that the machine should learn according to the base knowledge.

We are interested in Knowledge Guided Learning (KGL), the class of models where KENN belongs. The objective is to improve the performance of a NN model, by providing a Prior Knowledge which is expressed in terms of logical formulas and which acts as a supervisor for the learning process. At present, two main ways of injecting knowledge inside NN have been employed. The first one involves the usage of a regularization term in the loss function. Indeed, logical rules can be interpreted as constraints for the weights in the learning process; in ML, the natural way to introduce constraints in learning is to add a penalization term in the loss function, which represents in some way the satisfaction of the logical rules. The second approach, adopted by KENN, is to directly modify the network architecture: in this way, logical rules are enforced at the level of the network topology.

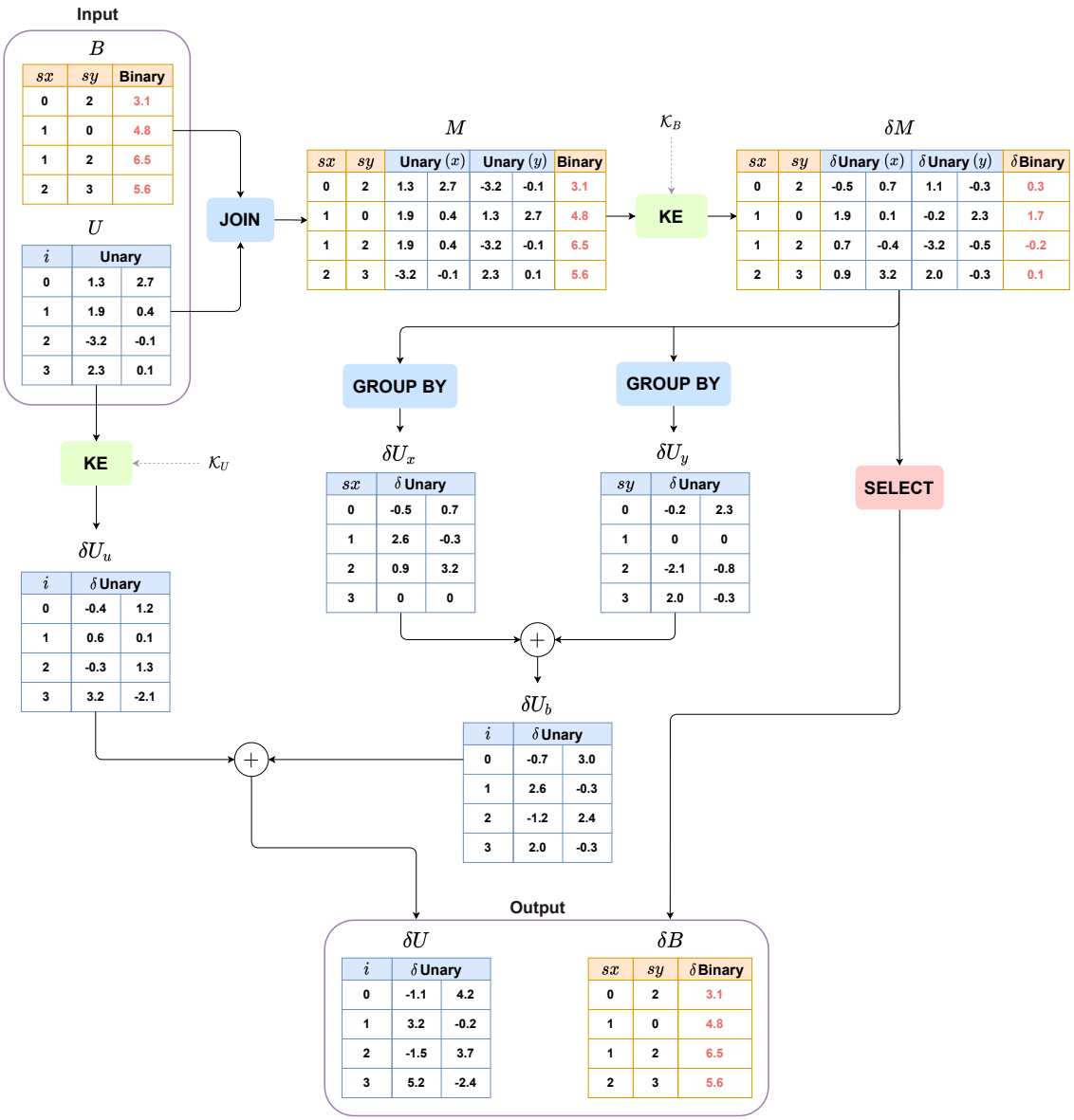


Figure 3.5: This figure shows an example with all the necessary computations to compute the final delta matrices δU and δB (in the bottom), starting from matrices U and B (top left).

3.4.1 Regularization Approaches

In KGL, regularization approaches enforce the satisfaction of the knowledge by defining a special penalization term to be applied to the loss function. Here we briefly mention two important examples of regularization based approaches: Logic Tensor Networks and Semantic Based Regularization.

Logic Tensor Networks

Logic Tensor Networks (LTN) [51] is a notable example of a method capable of integrating logical knowledge inside a NN by directly modifying the loss function. To do this, the authors define a differentiable first-order logic language called Real Logic, with which they manage to represent common deep learning tasks, such as clustering, multi-label classification, relational learning, regression, embedding learning and many others. In simple terms, the role of Real Logic is to act as a bridge between the purely symbolic world of logic, and the sub-symbolic world of neural systems. This is achieved by defining a specific semantic for the language, where each domain is interpreted as sets of tensors in the real field. In the same way, each constant, variable and term of the language is interpreted as a tensor of real values, function symbols are interpreted as functions between tensors, and predicates are interpreted as functions that map tensors into the interval $[0, 1]$. Specifically, given s any symbol of the language $\mathcal{L}$, the authors call its interpretation the "grounding" of s, and denote it with $\mathcal{G}(s)$[‡].

The authors then define how to compute the grounding of formulas, by using the semantics of first-order fuzzy logic. The way in which learning becomes possible is by defining the parametric grounding for symbols: given a symbol s, the parametric grounding of s is a grounding which is not known in advance, and can be computed exclusively by knowing a set of parameters. It is denoted as $\mathcal{G}(s|\theta_s)$, where θ_s is the set of parameter values that uniquely determines the value of the grounding. With this setup, learning is defined as the process of searching for the set of parameters θ^* such that:

$$\theta^* = \underset{\theta \in \Theta}{\operatorname{argmax}} \left(\underset{\phi \in \mathcal{K}}{\mathrm{SatAgg}} \mathcal{G}_\theta(\phi) \right),$$

where the quantity $\underset{\phi \in \mathcal{K}}{\mathrm{SatAgg}} \mathcal{G}_\theta(\phi)$ denotes the level of satisfiability of all the formulas in the

[‡]Note how here the term "grounding" differs from the standard meaning, which was defined in Section 3.1. While talking about LTN, the term "grounding" should be read as "interpretation".

knowledge base, with respect to a given aggregating operator $\mathrm{SatAgg} : [0,1]^* \to [0,1]$, which has the task of aggregating all the truth values of the formulas.

It's interesting to note that, based on what kind of symbol we are learning the grounding for, one can identify a corresponding task in machine learning:

- If $s \in \mathcal{C}$, corresponds to learning an embedding;
- If $s \in \mathcal{F}$, it corresponds to learning regression tasks;
- If $s \in \mathcal{P}$, it corresponds to learning a classification task.

For the simple case of binary classification, for example, the task is to learn the parametric grounding of a predicate A, $\mathcal{G}(A|\theta) : x \to \sigma(\mathrm{MLP}_\theta(x))$, where MLP denotes a multi layer perceptron. If we define with D the dataset with all the examples, the actual loss function to minimize will be:

$$L = \left(1 - \underset{\phi \in \mathcal{K}}{\mathrm{SatAgg}} \mathcal{G}_{\theta, x \leftarrow B}(\phi(x))\right),$$

where the notation $\underset{\phi \in \mathcal{K}}{\mathrm{SatAgg}} \mathcal{G}_{\theta, x \leftarrow B}(\phi(x))$ means that the variable x is grounded with the data B, where B is a mini-batch sampled from D.

Semantic Based Regularization

Semantic Based Regularization (SBR) is a general learning framework designed to integrate domain specific background knowledge in the form of first-order logic (FOL) clauses. To enforce the satisfaction of all the clauses, SBR introduces special regularization terms in the loss function, which represent the satisfaction of the knowledge. Specifically, given a background knowledge represented by set of H clauses, the satisfaction of the h-th clause can be represented by the quantity denoted by $0 \leq \phi_h(f) \leq 1$, where f is the vector of predictions provided by the model. From here, the regularization term to be added to the loss function is defined as

$$\sum_{h=1}^{H} \lambda_h (1 - \phi_h(f)),$$

where λ_h is the weight associated to the h-th constraint. A higher value of λ_h will increase the cost of not satisfying the constraint, meaning that the importance of the corresponding rule will increase. The conversion of FOL clauses into differentiable functions is made possible by considering fuzzy generalizations of FOL logic, similarly to how it is done in KENN. This is

a natural approach for machine learning tasks, but a major disadvantage over KENN is that, since the clause weights are introduced at the level of the loss function, those cannot be learned and are required to be known in advance. This, of course, is unlikely to happen in real scenarios and it is much more desirable to learn the weights together with the other learnable parameters of the model.

3.4.2 Model Based Approaches

An example of model based approach for the integration of logic in NNs is represented by Relational Neural Machines (RNM) [3]. A RNM models a probability distribution over a set of m output variables, given the predictions provided by one (or more) base NN, and a set of model parameters. Specifically, differently from a standard NN, each forward pass in RNM is performed in two separate stages: in the first one, the predictions from the NN are obtained from the input data; the second is a *semantic* stage, where the logical costraints are enforced over the output. More precisely, this second stage is performed by an undirected graphical model. RNM defines a conditional probability distribution of the exponential family over the output variables, which is defined as follows:

$$p(y|f,\lambda) = \frac{1}{Z}\exp\left(\sum_{x\in S}\Phi_0(f(x),y(x)) + \sum_c \lambda_c \Phi_c(y)\right),$$

where Z is the partition function, $f(x)$ are the predictions from the NN, λ is the set of parameters driving the semantic stage, Φ_0 is a potential function which enforces the supervised learning paradigm, S is the subset of supervised input vectors and Φ_c is a potential associated to c, a FOL formula part of the base knowledge. Specifically, Φ_c enforces the satisfaction of the clause c, while λ_c determines the weight of such clause.

The approach used by RNM is similar to the one used in KENN: a base NN (or any learner) is used as a foundation which provides initial predictions: on those, the model performs a post elaboration step, enforcing logical rules and improving the predictions. Also, like in KENN, the clause weights are trained together with the whole model, which is a desirable thing in real use cases. However, one significant drawback of RNM is that, for each training step, and also at inference time, an optimization problem must be solved. Specifically, the best assignment for the output variables is found with a Maximum a Posteriori (MAP) estimation, which maximizes the posterior probability of the grounded target variables, given the predictions from the

base NN and the parameters λ:

$$y^* = \underset{y}{\text{argmax}} P(y|f, \lambda).$$

In comparison, KENN is much more efficient since it is implemented as an actual layer of the NN architecture, which is then trainable end-to-end, together with the clause weights.

3.5 Experiments

In this section we describe in detail the experiments performed with KENN. We tested the ability of KENN of working with relational data, in the context of Collective Classification [52], both with the inductive and transductive learning paradigm.

Definition 3.5.1 (Collective Classification)**.** Consider a directed graph, consisting in a set of nodes V and a set of edges E. Each $v \in V$ is described with a vector of features $x \in \mathbb{R}^n$ and belongs to one of k classes $\{\omega_i\}_{i=1}^k$. The set of nodes V is further divided in two subsets of nodes: X, the set of nodes for which the correct label is known (Training Set), and Y, the set of nodes for which it is unknown (Test Set). The task of Collective Classification is to correctly predict the labels of nodes in Y, given the feature vectors of X and the topology structure determined by E.

Two learning paradigms can be defined:

- **Inductive Learning**: two separate graphs are used: $G_x = (X, E_x)$ for training and $G_y = (Y, E_y)$ for testing, where $E_x = \{(u, v)|u, v \in X\}$ and $E_y = \{(u, v)|u, v \in Y\}$. In other words, the edges of nodes between train and test set are not considered.
- **Transductive Learning**: a single graph is considered, with both training and test nodes: the network can use the information coming from the relations from the Test set even during training, but the actual supervision will come only from training nodes. Compared with the inductive paradigm, here the network has more available information, and for this reason we can expect better results.

We also provide a comparison of our results with the ones from SBR and RNM, reported on [3], on the same dataset and using the same learning paradigms. All the experiments were carried out with Python 3 and TensorFlow 2 [50]. All the code is publicly available on Github[§].

[§] `https://github.com/rmazzier/KENN-Citeseer-Experiments`

3.5.1 Citeseer Dataset

The experiments were conducted on the Citeseer Dataset [53]: it consists in a citation network with 4732 citations (directed links) between 3312 scientific publications (nodes), belonging to 6 different classes which represent the topic of the paper. Each node in the dataset is represented by a $0/1$ valued feature vector, where each entry indicates the absence or presence of the corresponding word in the dictionary, which is constituted by 3703 unique words.

3.5.2 The Prior Knowledge

The knowledge that we want to use in order to improve the predictions from the NN is the intuitive fact that, if a paper cites another paper, it is probably true that they are of the same topic. If we denote with $T_i(x)$ the truth value that node x belongs to the i-th output class, this fact can be encoded in terms of a logical clause as follows:

$$\forall x \forall y \quad T_i(x) \wedge \mathrm{Cite}(x, y) \rightarrow T_i(y), \quad i = 1, \ldots, 6$$

meaning that the this clause is repeated one time for each different output class. Inside KENN, this clause is represented as a disjunction of literals as follows:

$$\forall x \forall y \quad \neg T_i(x) \vee \neg\, \mathrm{Cite}(x, y) \vee T_i(y), \quad i = 1, \ldots, 6$$

3.5.3 Experimental Setup

The aim of these experiments was to obtain comparable results to those reported in [3]. For this reason we used the same base NN that was used there, which consists in a fully connected dense NN with three hidden layers, each of which using 50 hidden units and the ReLu activation function. Specifically, the NN takes in input only the features of each document in the dataset, meaning that it will be blind to the relations between them. For this reason, we can already observe that the predictions of the NN will not be influenced in any way by the choice of the learning paradigm. Specifically, we used the same random seed for both the tasks, so the the predictions of the NN are exactly the same between the two paradigms. The relational information (i.e. the citations between the papers) is introduced only at the level of the KE, and is treated as base knowledge. Note that this is a very specific case: KENN, in general, is designed to be able to learn both unary and binary predicates. In this case instead, the relational knowledge

is provided as ground truth and is used as an additional piece of information, which is exploited by the model to produce better predictions. Also note that the only learned predicates are the unary ones. Specifically, the forward pass is constituted by the following steps:

1. The base NN takes in input the matrix of features for the current batch of data, and returns the matrix of unary preactivations U (without the column containing the object identifiers, which is left implicit);

2. The matrix of binary predicates B is derived directly from the data, according to the current learning paradigm being considered. Since there is just a binary predicate, namely the Cite predicate, matrix B will have just three columns, two for the indices of the object pairs, and one for the truth value of the binary predicate.

3. KENN takes in input U and B, and produces deltas δU and δB, as described in Section 3.3. Since we are just interested in learning unary predicates, we just keep δU, while δB is discarded;

4. Differently from the standard case where the sigmoidal activation function is applied in the KE module, here the KE uses the softmax activation function. This is because for these experiments we are considering a multi-class classification task where each object can belong exclusively to one class. Specifically, the output of the KENN layer is obtained as follows:
$$y' = \mathrm{softmax}(U + \delta U).$$

In Figure 3.6 an example with all the listed steps is reported. Additionally, note that the truth values of the connections for this dataset are hard, meaning that nodes can be either connected or not connected, with no other alternative. For this reason pairs of connected objects are assigned a very high value, in this case 500[¶]. Also notice that only the pairs of connected objects are reported in B. In fact, it would be useless to consider also the pairs of not connected nodes (thus having a preactivation value equal to -500 for the Cite predicate), since for those cases the grounded clauses of our knowledge would be automatically satisfied[‖], and KENN would not apply any change. Thus, the number of rows of matrix B will be equal to the number of edges in the training graph.

In our model architecture, we also stack three KENN layers instead of only one: this choice is motivated by the fact that a single layer would consider only the neighbors of each node. By

[¶]There is not a specific reason for the choice of this number. The only important thing is that, when applying the sigmoidal activation function, the resulting truth value is close to 1.

[‖]Take c_1 and c_2 two nodes in the graph which are not connected; given the clause $\neg T(c_1) \vee \neg \mathrm{Cite}(c_1, c_2) \vee T(c_2)$, its truth value will be always ≈ 1, since $\mathcal{I}(\neg \mathrm{Cite}(c_1, c_2)) \approx 1$.

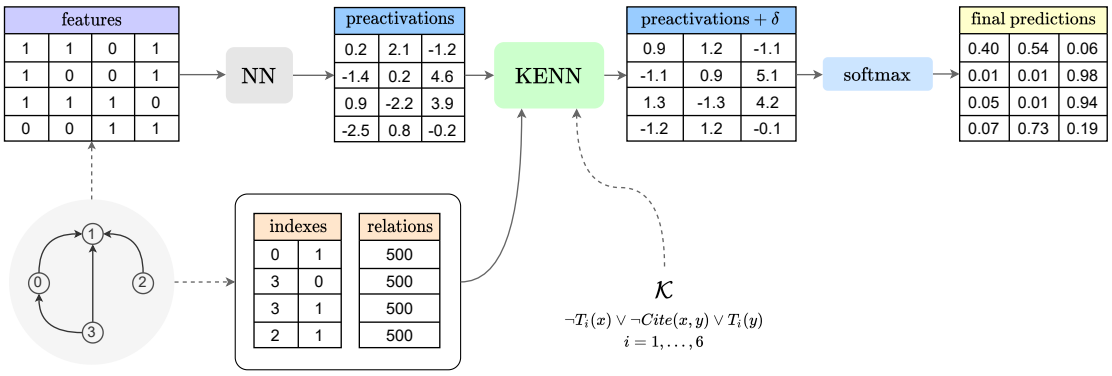


Figure 3.6: Simple example showing how the relational knowledge is injected in the NN for the Citeseer experiments. In this toy example, features are 4-dimensional vectors and there are 3 unary predicates; in the actual experiments, feature vectors have 3703 components and the output classes are 6.

adding three layers, instead, we allow for a propagation of the changes to farther nodes. This intuition was confirmed by empirically better results with respect to ones using a single KENN layer.

In addition to considering two different learning paradigms, we also evaluate our model on splits of different sizes, in order to see the impact of the dataset size on the final results. Specifically, for each learning paradigm, the model is trained on 10%, 25%, 50%, 75%, 90% splits of the whole dataset. Additionally, to have statistically relevant results, for each training percentage the dataset was trained for 500 times, where each time the training data is randomly picked in such a way that the class balance between train validation and test sets is preserved. Such a high number was motivated by the fact that, when training on always different splits, variance of the results can be very high, especially in the cases where the training set constitutes a small percentage of the whole dataset. In [3], the same procedure has been adopted, with the only difference that the number of training runs for each split percentage was 10 instead of 500. This was probably due to much longer training times: recall that in the case of RNM an optimization problem must be solved at each training step and also at inference time. Indeed, this fact highlights how one of the main advantages of KENN is its efficiency and scalability. Due to the different number of training runs, results won't be exactly comparable to the ones provided in [3]; nevertheless, our results will have a strong statistical significance. Additionally, we also provide 95% confidence intervals for each test set accuracy, as well as p-values with respect to the null hypothesis that the mean test set accuracy of KENN is the same as the one of the base NN.

3.5.4 Results

Inductive Learning

In Table 3.1 we report the results of the experiments for the inductive case. Specifically, each column contains the mean of all the test set accuracies from the different runs. The results for SBR and RNM are the ones reported in [3]. Also note that since we were not able to perfectly replicate the results for the base NN accuracies reported in [3], we report the ones obtained by our base NN, and use the relative improvements as the main evaluation metric. The relative improvements are also visualized in Figure 3.7, together with 95% confidence intervals, with a comparison with those from SBR and RNM. Finally in Figure 3.8, we report histograms showing the distribution of the test set accuracies of both our base NN and KENN, together with the distribution of the relative improvements for all the runs. The p-values for every training percentage were extremely small and well below the 0.05 threshold: for completeness, they are reported in Table 3.3. This means that, for both the training paradigms, we can safely reject the null hypothesis and be almost sure that the improvements provided by KENN are not a result of random perturbations due to the different choice of the training splits. As we can observe,

Table 3.1: Test set accuracies obtained with the inductive paradigm. The columns for SBR and RNM show the results reported in [3]. The quantities between parentheses denote the relative improvement with respect to the base NN.

% training	NN	SBR	RNM	NN	KENN
10	0.645	0.650 (+0.005)	**0.685** (+0.040)	0.544	0.601 (+**0.048**)
25	0.674	0.682 (+0.008)	**0.709** (+0.035)	0.629	0.671 (+**0.041**)
50	0.707	0.712 (+0.005)	**0.726** (+0.019)	0.680	0.714 (+**0.034**)
75	0.717	0.719 (+0.002)	0.726 (+0.009)	0.733	**0.754** (+**0.021**)
90	0.723	0.726 (+0.003)	0.732 (+0.009)	0.759	**0.768** (+**0.010**)

KENN outperforms both RNM and SBR for each training percentage. Additionally, we can see how the relative improvement becomes smaller as the training percentage increases. This makes sense from an intuitive point of view: when disposing of few data, NNs struggle to give good results and the prior knowledge plays a more important role, providing more improvements. On the other hand, when the NN is trained on a lot of data, the knowledge becomes

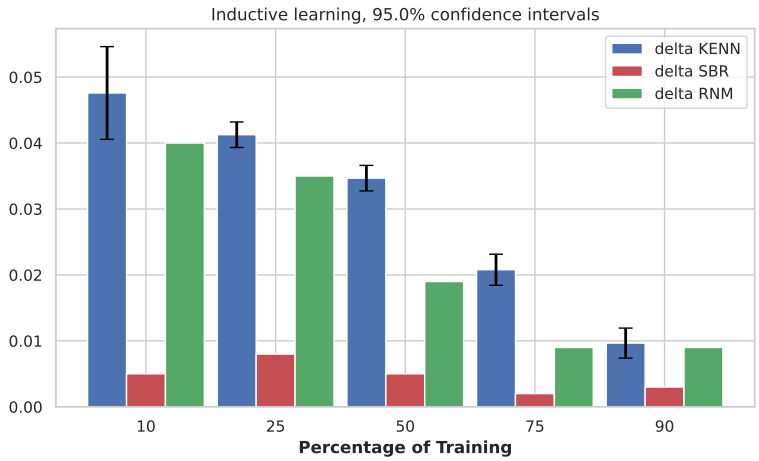


Figure 3.7: Relative improvements for the inductive learning task. 95% confidence intervals are provided for our results.

less important: indeed the NN has now access to much more information and, in some sense, can learn the rules on its own from the patterns in the data. In this sense, RNM shows a similar behavior with respect to KENN, even if providing less relative improvement overall.

Transductive Learning

We report the same results for the transductive case in Table 3.2, Figure 3.9 and Figure 3.10 respectively. Note that the results of our NN are exactly same of the ones from the inductive paradigm; this is because we used the same random seed for both the paradigms, and, as already mentioned, the base NN is blind to the learning paradigm. On the other hand, results of the NN from [3] report different results. Looking at the relative improvements, we can observe how all the three methods provide similar results for all the training dimensions, the only exception being the 10% training percentage, where KENN shows a much higher relative improvement. KENN still seem to outperform SBR and RNM for the $10\%, 25\%$ and 50% training percentages, but the difference is far less pronounced. However, recalling how the results from [3] are the average of just 10 runs, the comparison cannot be considered very reliable.

3.5.5 Clause Weights and satisfaction of the rules

One of the best features of KENN is its capability to learn the clause weights. This is a fundamental property, since it allows the model to automatically boost the importance of useful

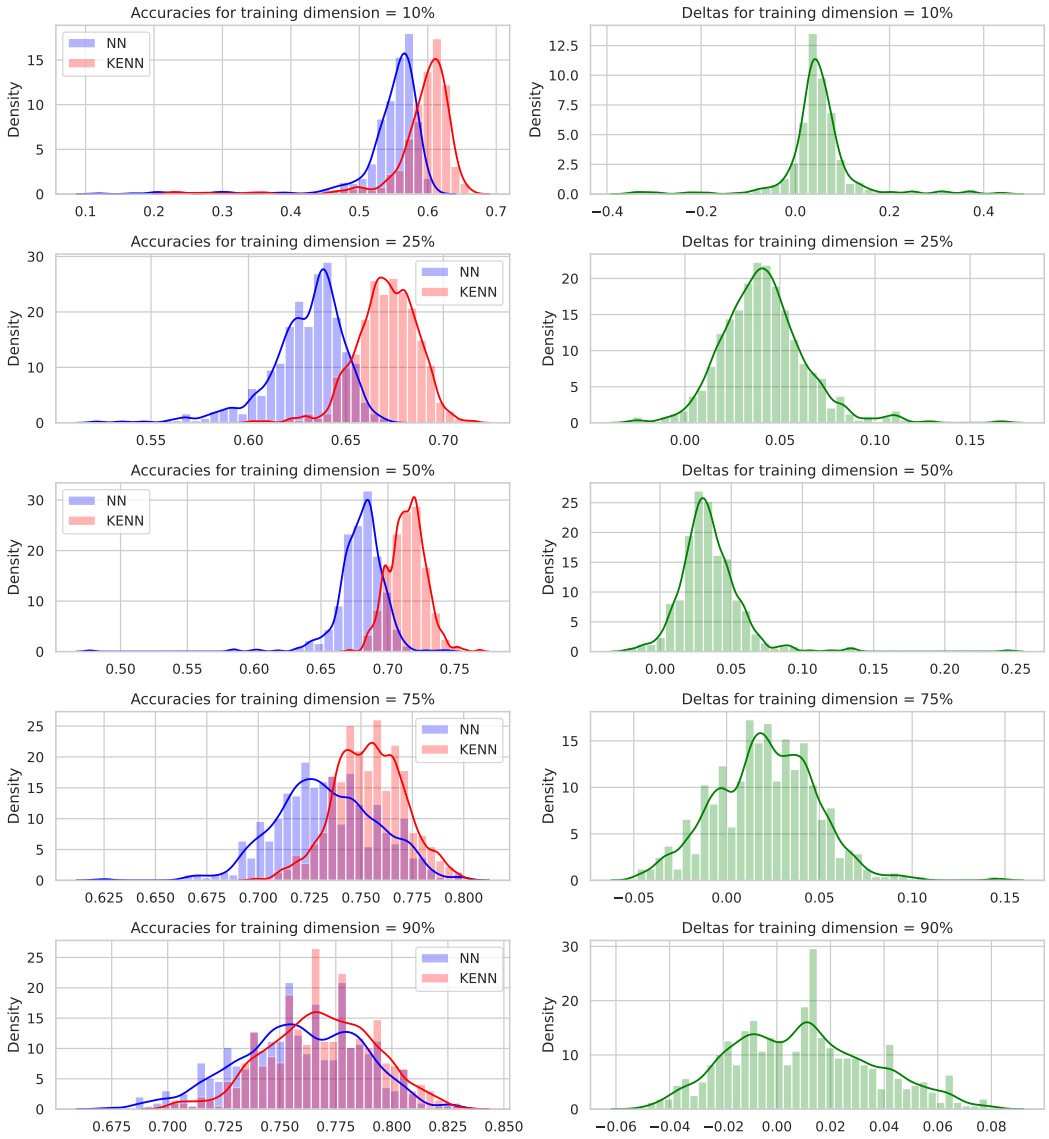


Figure 3.8: Histograms showing the distribution of the accuracies for all the different 500 runs, for the inductive case. On the left, the accuracies of the base NN vs accuracies of KENN. On the right the distribution of the relative improvements.

Table 3.2: Test set accuracies obtained with the transductive paradigm. The columns for SBR and RNM show the results reported in [3]. The quantities between parentheses denote the relative improvement with respect to the base NN.

% training	NN	SBR	RNM	NN	KENN
10	0.640	0.703 (+0.063)	**0.708** (+0.068)	0.544	0.652 (+**0.108**)
25	0.667	0.729 (+0.062)	**0.735** (+0.068)	0.629	0.702 (+**0.073**)
50	0.695	0.747 (+0.052)	**0.753** +0.058)	0.680	0.744 (+**0.065**)
75	0.708	0.764 (+0.056)	0.766 (+**0.058**)	0.733	**0.788** (+0.055)
90	0.726	0.780 (+**0.054**)	0.780 (+**0.054**)	0.759	**0.808** (+0.049)

Table 3.3: p-values computed for each learning paradigm and for each training percentage. Values below 10^{-100} are reported as 0.

% training	p-value Inductive	p-value Transductive
10	$3.22 \cdot 10^{-36}$	0
25	0	0
50	0	0
75	$2.25 \cdot 10^{-48}$	0
90	$2.87 \cdot 10^{-9}$	0

rules, and to ignore useless ones. However, to actually verify that KENN is able to correctly learn these weights, we analyzed them after training and investigated on the reasons that lead to their growth or shrinkage.

Specifically, what we care about is to have an empirical evidence that, if a certain clause is particularly important, KENN is able to increase its respective clause weight. We quantify the importance of a specific clause by defining the *clause compliance*, a metric which expresses how much a clause is satisfied on the training data.

Definition 3.5.2 (Clause Compliance)**.** Let $G := (V, E, T)$ be the current training graph, where V is the set of nodes, E is the set of edges and $T : V \to \{1, \ldots, 6\}$ is the function that maps each node to its ground truth class. Given one of the output classes $k \in \{1, \ldots, 6\}$ and the corresponding clause $c_k : \neg k(x) \vee \neg \mathrm{Cite}(x, y) \vee k(y)$, we define the clause compliance

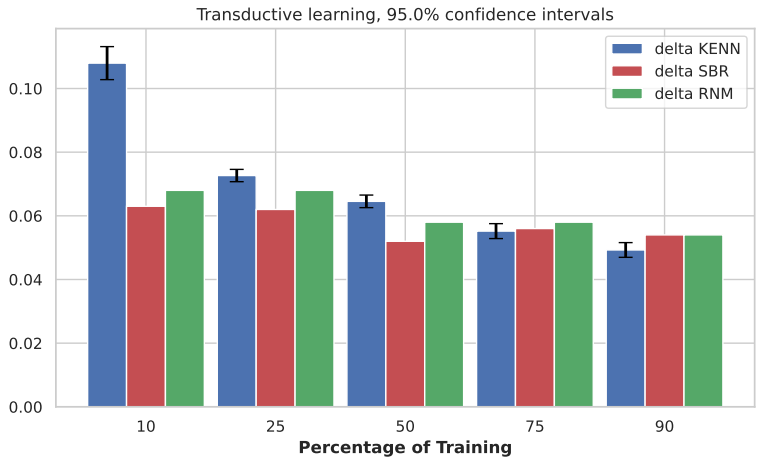


Figure 3.9: Relative improvements for the transductive learning task. 95% confidence intervals are provided for our results.

of c_k in G as:

$$C(G, c_k) = \frac{\sum_{v \in T_k} \sum_{u \in \mathcal{N}(v)} \mathbf{1}(u \in T_k)}{\sum_{v \in T_k} |\mathcal{N}(v)|} \tag{3.15}$$

where T_k is the set of nodes of topic k, $\mathcal{N}(v)$ is the number of nodes cited by v, and $\mathbf{1}(u \in T_k)$ is equal to 1 if $u \in T_k$ or 0 otherwise. In simpler terms, $C(G, c_k)$ is the ratio between the number of citations from papers of topic k to papers of topic k and the total number of citations coming from papers of topic k. It is equal to 1 when always satisfied, and is equal to 0 when never satisfied. To investigate whether KENN is capable to associate higher weights to clauses with a high clause compliance, we perform 85 different training runs, and inspect the learned weights for each clause against its clause compliance. We repeat this process for each percentage of the training set. Results are visualized in Figure 3.11: by looking at the obtained plots, we can observe that there is a strong correlation between the two variables, which gets more pronounced as the size of the training set increases. Another interesting fact is that, as the clause compliance decreases, an higher variance in the values of the clause weights is observed. For example, looking at the 90% case, we can see that the clause associated to the topic "AI" has a mean clause compliance slightly smaller than 0.5, meaning that this rule is satisfied slightly less than half of the times. By looking at all the values assumed by its clause weight throughout the runs, we can indeed see that the results are very variable with respect to the other topics. This behavior goes against our intuition in some way: if a rule is not useful (like in this case), then the clause weights should monotonically decrease during training, leading to low and tightly

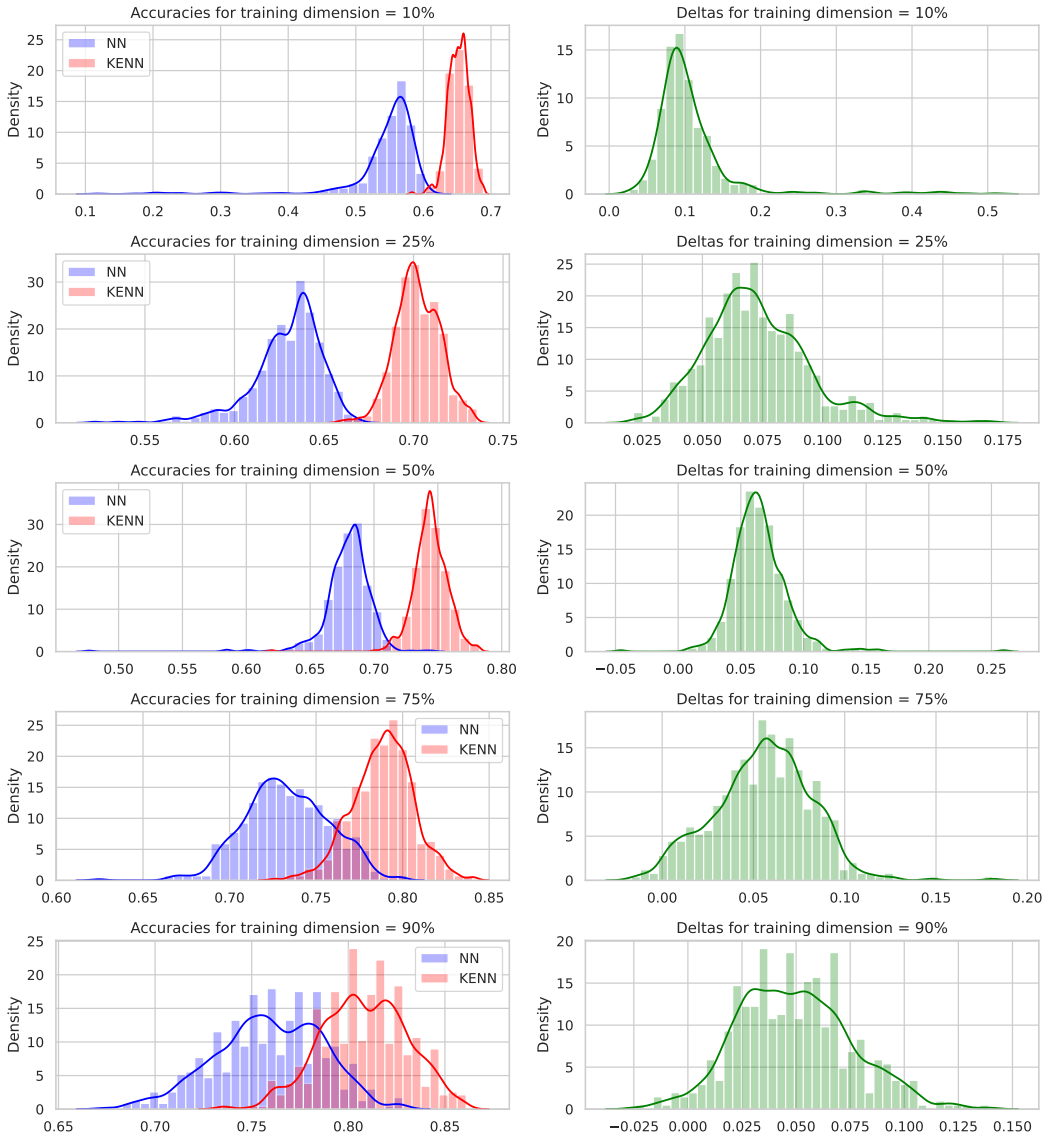


Figure 3.10: Histograms showing the distribution of the accuracies for all the different 500 runs, for the transductive case. On the left, the accuracies of the base NN vs accuracies of KENN. On the right the distribution of the relative improvements.

packed values for the learned clause weights. In reality, we found that as the compliance of the rules in the data decreases, the learning of their clause weights becomes more randomic: this could be due to a variety of factors, which do not appear to have an obvious or easy explanation. One possible interpretation could be that different randomic initializations of the base NN could lead less compliant clauses to actually produce positive changes, in turn leading KENN to believe that the importance of such a clause should be boosted.

3.6 Explainability in KENN

In this section, we study how to obtain explanations from KENN. In Chapter 2 we studied different kinds of approaches for interpretability and explainability in ML, specifically for the case of NNs. We also analyzed the important distinction between transparent models and post-hoc explainability methods, together with their advantages and disadvantages. As we know, KENN can't work on its own: it needs a base NN classifier which has the task of providing the initial predictions. For this reason, KENN can't be considered a completely transparent model just because it is designed to work alongside a standard NN architecture, which is inherently hard to interpret and explain. However, once the initial predictions from the base NN are provided, everything that happens inside the KE can be interpreted; for this reason, KENN can be considered a partially transparent model. Indeed, if we restrict just to the phase where the rules from the knowledge are enforced, there is a straightforward way to visualize how KENN modified the predictions of the NN, that is to inspect the delta vectors. Specifically, this process can happen at different levels of precision: for example, at inference time, we might want to study the changes caused by the whole base knowledge, or we might desire to isolate the effects of a single clause on a subset of input samples. More precisely, given a base knowledge $\mathcal{K}$, and given a subset of clauses $\mathcal{C} \subset \mathcal{K}$ for which we want to observe the effects, the delta vector we are looking for is the sum of all the vectors $\delta_c, \forall c \in \mathcal{C}$. However, note that the KE module is designed to compute $\sum_{\mathcal{K}} \delta^c$, while we are interested in $\sum_{\mathcal{C}} \delta^c$: for this reason we will need to intercept the deltas just as they are returned by their respective CEs. Figure 3.12 provides a visualization of this process for the unary and binary case. We can also extract the deltas caused by different non overlapping subsets of clauses. Note however that the exact contribution from each different subset can be visualized only when working at the level of preactivations. Indeed, we know that the actual delta at the activation level (denoted in Section 3.1.4 as δ^g) will change based on the preactivations values, even if the initial delta applied at the preactivation level (de-

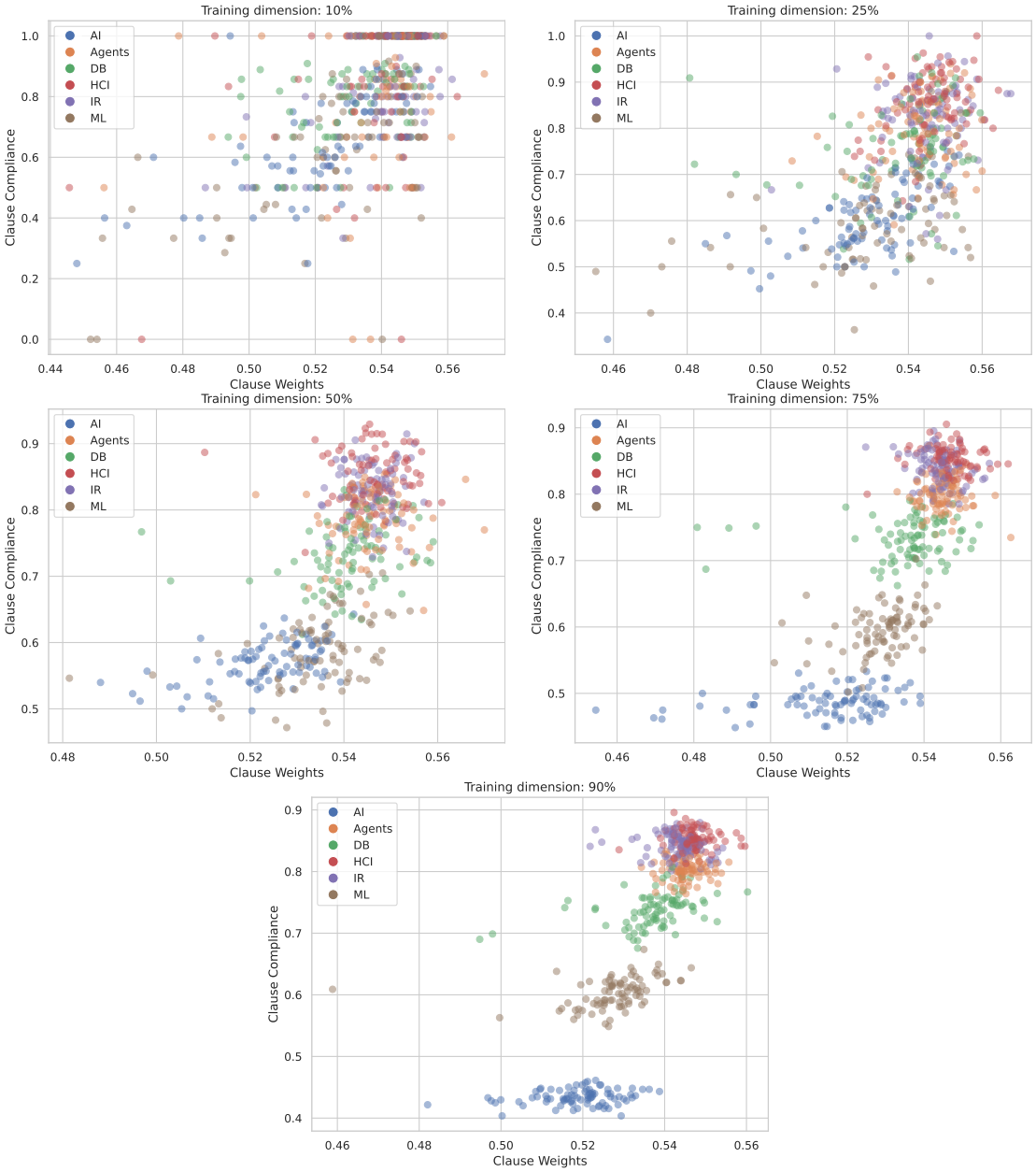
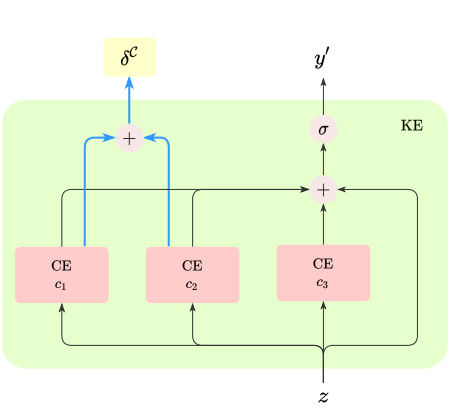
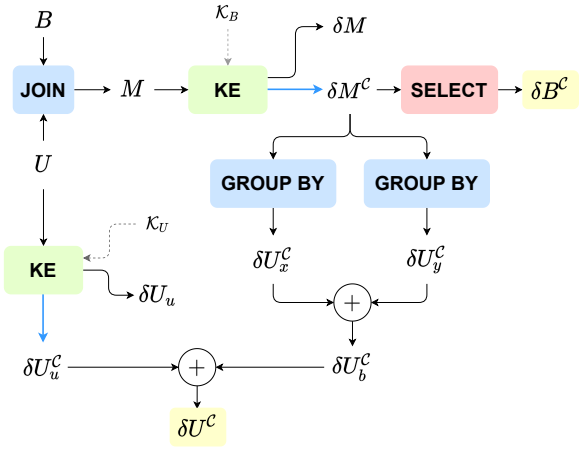


Figure 3.11: Scatterplots showing the relation between clause weight and clause compliance, for each clause from 85 different runs, for each different training percentage. We can observe how, as the training dimension increases, KENN learns to adjust the clause weights according on how much that clause is satisfied in the training set. Each dot in the scatterplots corresponds to a clause in a specific run; the colour of the dot denotes the topic related to that clause.



(a) Delta extraction from the KE module, for the clauses subset $\mathcal{C} = \{c_1, c_2\}$.



(b) Delta extraction for the binary case. Blue arrows denote the extraction operation depicted on Figure 3.12a.

Figure 3.12: Process for extracting deltas relative to the clauses in $\mathcal{C} \subset \mathcal{K}$. The output deltas are highlighted in yellow for both the unary case (a) and binary case (b).

noted in previous section as δ^f) is the same. Recall that we already illustrated this fact in Figure 3.1. This implies that, when cumulating more deltas from different clauses, we can visualize the exact contribution from each clause at the preactivation level, but we lose this information at the activation level, where the only thing we can observe is the aggregated contribution of all the clauses. In other words, the contribution of each clause at the activation level depends on the order in which the clauses apply their contribution at the preactivation level, which is not defined since all the deltas are applied simultaneously. A simple example showing this problem is reported in Figure 3.13.

Another nice property of KENN is the possibility to quantify at inference time how much the contributions from the Base Knowledge improved (or worsened) the predictions of the base NN for the current sample being classified. For example, we can define the following metric:

Definition 3.6.1 (Improvement Score)**.** Given an input sample x to be classified into m different classes, given the vector of deltas $\delta^{\mathcal{C}} \in \mathbb{R}^m$ coming from clauses belonging to $\mathcal{C} \subset \mathcal{K}$ and given vector of ground truth labels $l \in \{-1, 1\}^m$, where entries corresponding to the positive class are equal to 1 and -1 otherwise, then the improvement score for sample x with respect to $\mathcal{C}$ is defined as follows:

$$IS(x, \mathcal{C}) = \sum_{i=1}^{m} -\delta_i^{\mathcal{C}} \cdot l_i = -\delta^{\mathcal{C}\top} \cdot l. \tag{3.16}$$

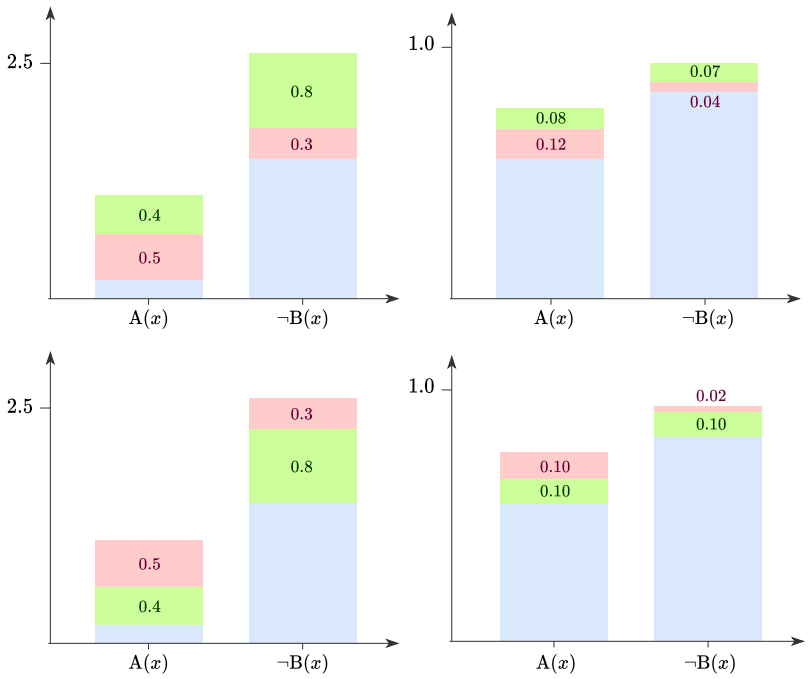


Figure 3.13: The Figure shows an example where we apply deltas for two different clauses, in different orders, at the pre-activation level (left), and the resulting deltas at the activation level (right). The deltas from each clause are represented by different colors, while the predictions from the base NN are shown in blue. On the right, the deltas at the activation level are shown: note how, when changing the order of application of the same deltas at the preactivation level, the deltas at the activation level change.

Specifically, the improvement score quantifies the positive contribution of the KENN layer for the current prediction.

In simple terms, what happens in equation (3.16) is that each term of the delta vector will produce a positive contribution only when its sign is the same as the corresponding ground truth label.

Additionally, recall that KENN aggregates the deltas of all the clauses by simply summing the individual deltas: this is very convenient since it makes KENN fast and scalable. However, we saw that this kind of aggregation may be too simplistic, and may cause inconsistencies in the case where multiple clauses disagree on how certain truth values should be changed. In order to diagnose this kind of behavior on a trained KENN model, we can use the following score.

Definition 3.6.2 (Disagreement Score)**.** Using the previously defined notation, the disagreement score for sample x with respect to the subset of clauses $\mathcal{C}$ is defined as follows:

$$DS(x, \mathcal{C}) = \sum_{c \in \mathcal{C}} |\delta_c| - \left| \sum_{c \in \mathcal{C}} \delta_c \right|. \tag{3.17}$$

This score quantifies the amount of inconsistencies and disagreement among the clauses belonging to $\mathcal{C}$: a score of 0 implies perfect agreement between all the clauses, while an higher score reflects an higher disagreement.

These defined metrics can be useful for different situations, like during debugging or for the model evaluation phase. For example, we might want to analyze the predictions for which KENN gave the most improvements or, viceversa, we might want to see where KENN has actually provided worse results with respect to the base NN. In general, explanations extracted from KENN can be useful mostly to evaluate the impact of the knowledge, and can allow the user to refine it, by adding or removing rules that proved to be useful or damaging. We remark, however, that this transparency is just partial and limited to the knowledge enforcement stage: indeed, the base NN remains an opaque component of the whole model, even if used in conjunction with KENN.

4

Conclusion

In this work we first gave a review of XAI methods, and distinguished two main main approaches for explainability. The first is transparency, which refers to the particular property of models to be interpreted by human supervisors without the need of external tools or complex post-elaboration methods. Transparency is desirable since allows us to have a precise idea of how predictions are obtained, by knowing the internal mechanisms of the models. One drawback, however, is that, tipically, for a model to be transparent it is also required to be relatively simple and to have few parameters; this goes against the latest finding in DL research, where deeper models seem to be the more performing ones. On the other hand, post-hoc explainability focuses on extracting explanations from any kind of trained model, which is free to retain its black box nature. These approaches are more complex and often require to solve additional optimization problems, but their ability to be applied to any trained model makes the performance sacrifice no longer necessary.

We then described KENN, a special residual layer designed to inject prior knowledge, expressed as a set FOL clauses, inside a base NN classifier. Specifically, its objective is to improve the base NN predictions by maximizing the truth values of all the rules in the knowledge base. This is made possible thanks to TBFs, which, applied to preactivations from the NN, increase the satisfaction of the rules with respect to the training data. We also compared KENN with SBR and RNM, two other notable examples of neuro-symbolic architectures, and analyzed their differences, advantages and shortcomings. A remarkable feature of KENN is its capability to learn the clause weights, parameters associated to each logical rule, which define their impor-

tance. We explored this specific feature of KENN by visualizing the relationship between the clause compliances in the training data with the corresponding clause weights, and noticed a linear correlation which gets stronger as the amount of training data increases. This is an evidence of how KENN is able to boost the importance of useful rules, while partially ignoring useless or damaging ones. We also provided experimental results on the task of collective classification on the Citeseer Dataset: specifically, for the inductive learning paradigm, we found that KENN outperforms both SBR and RNM. Finally, we also propose methods to extract explanations from KENN. Specifically, we saw how to inspect the precise changes caused by any subset of rules from the knowledge base, by simply extracting the delta vectors from their respective CEs. We also propose some examples of custom metrics, which allow us to quantify the the positive (or negative) contribution of KENN in the context of local explanations, and also to assess the amount of disagreement between rules in the knowledge base.

Concluding, KENN is a promising technique to integrate prior knowledge inside NNs: differently from other NeSy methods, its is simple and computationally efficient, despite being less general and precise for other aspects (think for example to the way KENN aggregates the deltas, or to the possibility to use only the universal quantifier). The ability of KENN to learn the clause weights can be an interesting subject of further research. For example, KENN could be used as a knowledge extraction method from data: starting from randomly generated rules, KENN can be able to selectively boost the weights of the more useful ones. Also, more experiments on simple unary base knowledge can be made. Finally, explainability in KENN is also a promising research topic: for example, in my experience of stage, we started working on an interactive, web-based interface to debug and visualize the results of a KENN model. This kind of tools can be very useful: they can encourage a new approach towards a more explainable AI, and boost popularity of neuro-symbolic models like KENN.

References

[1] A. Daniele and L. Serafini, "Knowledge enhanced neural networks," in *PRICAI 2019: Trends in Artificial Intelligence*. Cham: Springer International Publishing, 2019, pp. 542–554.

[2] A. Mordvintsev, C. Olah, and M. Tyka, "Inceptionism: Going deeper into neural networks," 2015. [Online]. Available: https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

[3] G. Marra, M. Diligenti, F. Giannini, M. Gori, and M. Maggini, "Relational neural machines," *arXiv preprint arXiv:2002.02193*, 2020.

[4] G. Marcus, "Deep learning: A critical appraisal," Jan. 2018.

[5] C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown, 2016.

[6] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko *et al.*, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[7] T. Gebru, J. Krause, Y. Wang, D. Chen, J. Deng, E. L. Aiden, and L. Fei-Fei, "Using deep learning and google street view to estimate the demographic makeup of the us," *arXiv preprint arXiv:1702.06683*, 2017.

[8] J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel *et al.*, "Mastering atari, go, chess and shogi by planning with a learned model," *Nature*, vol. 588, no. 7839, pp. 604–609, 2020.

[9] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[10] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain." *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[11] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," *arXiv preprint arXiv:1710.09829*, 2017.

[12] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," 12 2013.

[13] C. Buckner, "Adversarial examples and the deeper riddle of induction: The need for a theory of artifacts in deep learning," *arXiv preprint arXiv:2003.11917*, 2020.

[14] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," 06 2015, pp. 427–436.

[15] R. Jia and P. Liang, "Adversarial examples for evaluating reading comprehension systems," 01 2017, pp. 2021–2031.

[16] W. E. Zhang, Q. Z. Sheng, A. A. F. Alhazmi, and C. Li, "Generating textual adversarial examples for deep learning models: A survey," *arXiv preprint arXiv:1901.06796*, p. 129, 2019.

[17] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16. Red Hook, NY, USA: Curran Associates Inc., 2016, p. 4356–4364.

[18] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," *AI magazine*, vol. 38, no. 3, pp. 50–57, 2017.

[19] ACM. (2017) Statement on algorithmic transparency and accountability. [Online]. Available: https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf

[20] D. Gunning and D. Aha, "Darpa's explainable artificial intelligence (xai) program," *AI Magazine*, vol. 40, no. 2, pp. 44–58, Jun. 2019. [Online]. Available: https://ojs.aaai.org/index.php/aimagazine/article/view/2850

[21] S. Bromberger, *On what we know we don't know: Explanation, theory, linguistics, and how questions shape them*. University of Chicago Press, 1992.

[22] Z. Lipton, "The mythos of model interpretability," *Communications of the ACM*, vol. 61, 10 2016.

[23] F. Doshi-Velez and B. Kim. (2017) Towards a rigorous science of interpretable machine learning. [Online]. Available: https://arxiv.org/abs/1702.08608

[24] B. Kim, "Interactive and interpretable machine learning models for human machine collaboration," 2015.

[25] M. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," 02 2016, pp. 97–101.

[26] N. Burkart and M. Huber, "A survey on the explainability of supervised machine learning," *Journal of Artificial Intelligence Research*, vol. 70, 01 2021.

[27] B. Chen, Y. Li, S. Zhang, H. Lian, and T. He, "A deep learning method for judicial decision support," in *2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C)*, 2019, pp. 145–149.

[28] Y. Zhang and X. Chen, "Explainable recommendation: A survey and new perspectives," *arXiv preprint arXiv:1804.11192*, 2018.

[29] L. Gilpin, D. Bau, B. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," Oct. 2018, pp. 80–89.

[30] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1–15, 2018.

[31] B. Herman, "The promise and peril of human evaluation for model interpretability," *ArXiv*, vol. abs/1711.07414, 2017.

[32] F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: A survey," in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2018, pp. 0210–0215.

[33] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html

[34] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," 11 2014.

[35] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *preprint*, 12 2013.

[36] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, "Dueling network architectures for deep reinforcement learning," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 1995–2003.

[37] A. S. Ross, M. C. Hughes, and F. Doshi-Velez, "Right for the right reasons: Training differentiable models by constraining their explanations," *arXiv preprint arXiv:1703.03717*, 2017.

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[39] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[40] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[41] D. H. Park, L. A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, and M. Rohrbach, "Multimodal explanations: Justifying decisions and pointing to the evidence," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8779–8788.

[42] A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra, "Human attention in visual question answering: Do humans and deep networks look at the same regions?" *Computer Vision and Image Understanding*, vol. 163, pp. 90–100, 2017.

[43] R. Caruana, H. Kangarloo, J. D. Dionisio, U. Sinha, and D. Johnson, "Case-based explanation of non-case-based learning methods." in *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 1999, p. 212.

[44] L. G. Valiant, "Three problems in computer science," *J. ACM*, vol. 50, no. 1, p. 96–99, Jan. 2003. [Online]. Available: https://doi.org/10.1145/602382.602410

[45] M. Richardson and P. Domingos, "Markov logic networks," *Machine learning*, vol. 62, no. 1-2, pp. 107–136, 2006.

[46] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

[47] J. Wang and P. M. Domingos, "Hybrid markov logic networks." in *AAAI*, vol. 8, 2008, pp. 1106–1111.

[48] T. R. Besold, A. Garcez, S. Bader, H. Bowman, P. M. Domingos, P. Hitzler, K.-U. Kühnberger, L. Lamb, D. Lowd, P. Lima, L. Penning, G. Pinkas, H. Poon, and G. Zaverucha, "Neural-symbolic learning and reasoning: A survey and interpretation," *ArXiv*, vol. abs/1711.03902, 2017.

[49] V. Novák, "First-order fuzzy logic," *Studia logica*, vol. 46, no. 1, pp. 87–109, 1987.

[50] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, 2016, pp. 265–283.

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

[52] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad, "Collective classification in network data," *AI magazine*, vol. 29, no. 3, pp. 93–93, 2008.

[53] Q. Lu and L. Getoor, "Link-based text classification," in *IJCAI Workshop on Text Mining and Link Analysis*, 2003.

Acknowledgments

First of all, I would like to thank Alessandro Daniele, who followed and helped me both during and after my internship in FBK, proving to be an excellent and caring supervisor. I would also like to thank professor Luciano Serafini and all the members of the DKM research group at FBK, for providing a stimulating work environment and for always being ready to lend an hand in case of need.

I would like to thank my university colleagues and friends: in these two years we shared challenging moments and came out stronger together. This work was partly possible thanks to them and their help.

I would also like to thank my family, for their ever-present, unconditional and loving support. I could not be here without them.

Last but not least, I would like to thank Benedetta for her continuous, patient and loving support, which proved to be a constant source of strength.