PROTOtypical Logic Tensor Networks (PROTO-LTN) for Zero Shot Learning

PROTOtypical Logic Tensor Networks (PROTO-LTN) for Zero Shot Learning

Simone Martone, Francesco Manigrasso, Fabrizio Lamberti and Lia Morra
Department of Control and Computer Engineering
Politecnico di Torino
simone.martone@studenti.polito.it, {francesco.manigrasso, fabrizio.lamberti,lia.morra}@polito.it

***Abstract*—Semantic image interpretation can vastly benefit from approaches that combine sub-symbolic distributed representation learning with the capability to reason at a higher level of abstraction. Logic Tensor Networks (LTNs) are a class of neuro-symbolic systems based on a differentiable, first-order logic grounded into a deep neural network. LTNs replace the classical concept of training set with a knowledge base of fuzzy logical axioms. By defining a set of differentiable operators to approximate the role of connectives, predicates, functions and quantifiers, a loss function is automatically specified so that LTNs can learn to satisfy the knowledge base. We focus here on the subsumption or `isOfClass` predicate, which is fundamental to encode most semantic image interpretation tasks. Unlike conventional LTNs, which rely on a separate predicate for each class (e.g., dog, cat), each with its own set of learnable weights, we propose a common `isOfClass` predicate, whose level of truth is a function of the distance between an object embedding and the corresponding class prototype. The PROTOtypical Logic Tensor Networks (PROTO-LTN) extend the current formulation by grounding abstract concepts as parametrized class prototypes in a high-dimensional embedding space, while reducing the number of parameters required to ground the knowledge base.**

We show how this architecture can be effectively trained in the few and zero-shot learning scenarios. Experiments on Generalized Zero Shot Learning benchmarks validate the proposed implementation as a competitive alternative to traditional embedding-based approaches. The proposed formulation opens up new opportunities in zero shot learning settings, as the LTN formalism allows to integrate background knowledge in the form of logical axioms to compensate for the lack of labelled examples. PROTO-LTN was implemented in Tensorflow and is available at https://github.com/FrancescoManigrass/PROTO-LTN.git

I. Introduction

Despite their impressive performance when trained on large-scale, supervised datasets, deep neural networks have still difficulties generalizing to unseen categories. On the contrary, humans can leverage logical reasoning to make guesses about new circumstances, and are able to infer knowledge from few to zero examples. Recent efforts towards Neural-Symbolic (NeSy) integration [1], [2] allow to assimilate symbolic representation and reasoning into deep architectures: this entails that background knowledge, in the form of logical axioms, can be exploited during training, opening up new scenarios for settings in which labelled examples are scarce or noisy [3], [4]. Specifically, we focus here on Logic Tensor Networks (LTNs) [5], a NeSy architecture that replaces the classical concept of a training set with a Knowledge Base $\mathcal{K}$ of logical axioms, ultimately interpreted in a fuzzy way, and formulates the learning objective as maximizing the satisfiability of $\mathcal{K}$. While this framework has been applied to multi-label classification problems [5], [6] and object detection [4], its application to few- and zero-shot image classification has not yet been investigated.

In this work, we explore this task from a NeSy perspective, and propose to integrate ideas and concepts from the few-shot learning (FSL) and zero-shot learning (ZSL) domains, namely the Prototypical Networks (PNs) [7] framework, within the LTN formulation. PNs define class prototypes in a high-dimensional embedding space, so that incoming examples are assigned to the class of their nearest prototype according to some distance measure. In the LTN framework, this is achieved by representing the `isOfClass` relationship as a function of the distance between a class prototype and an object instance, thus obtaining the Prototypical Logic Tensor Network (PROTO-LTN) architecture. As the embedding space is the focus of the learning procedure, such prototypes may be also defined for classes that are not seen at training time.

The present study thus formulates a theoretical framework that achieves competitive results with respect to standard embedding-based ZSL architectures such as DEM [8], yet offering higher degrees of flexibility. Although our analysis shows that their basic settings the two formulations are equivalent, PROTO-LTNs have greater potential in both standard and transductive ZSL. They are able to integrate in the training process prior knowledge and logical constraints from an external knowledge base, including information related to unseen classes [9]. Hence, a NeSy formulation allows to constraint the embedding space via symbolic priors.

The proposed framework has also potential advantages over traditional LTNs, even outside of the FSL and ZSL settings, since classes are represented as parametrized prototypes rather than a discrete label space [5], [4]. First, representing higher-level concepts as distributed vectorized representations allows to naturally exploit the notion of distance for highlighting relationships between symbols, with semantically related symbols having similar representations [10]. Second, prototypes allow to ground abstract concepts in a vectorized form that can be more easily manipulated: as an example, it would be

easier to define a suitable grounding for predicates that directly operate on the abstract classes, as well as their instances. Third, prototypes are more interpretable than simple labels, as their incorporation into the embedding space can be easily visualized by employing dimensionality reduction methods, such as t-SNE [11].

The rest of the paper is organized as follows. In Section II, we place the present work in the context of the related literature, and provide a background on LTNs. In Section III, we describe a simple theoretical scheme to assimilate PNs into a LTN for classification purposes (PROTO-LTN), both in the FSL and ZSL scenarios. Then, in Sections IV and V, we examine the behavior of the model in the Generalized Zero-Shot-Learning (GZSL) task on common benchmark datasets. Finally, in Section VI, we discuss conclusions and future works.

II. Related work

A. Neural-symbolic AI in Semantic Image Interpretation

Research on how to combine connectionist and symbolic approaches has flourished in the past few years [5], [12], with several applications in semantic image interpretation and visual query answering [5], [4], [13], [3], [14], [15], [16]. Among the plethora of compositional patterns that have been proposed [17], [12], the present work follows two main principles: knowledge representation (in the form of first order logic) is embedded into a neural network, which in turn allows to constrain the search space by leveraging explicit (and human-interpretable) domain knowledge as a symbolic prior. This latter property is extremely useful in ZSL, in which some external source of information is exploited to offer an abstract description of the classes in lieu of providing training examples. On the other hand, compared to approaches based on Inductive Logic Programming (such as [14]), in which perception and reasoning are performed by separate modules, LTNs provide tighter integration between the two subsystems.

B. Logic Tensor Networks

LTNs have proven effective in higher-level image interpretation tasks, such as object detection and scene graph construction [13], [5]. Donadello et al. applied them for scene relationship detection in a zero shot setting, showing how prior knowledge can compensate for the lack of supervision [3].

In the LTN framework, the term *grounding* denotes the interpretation of a First Order Language into a subset of the $\mathbb{R}^n$ domain [5]. It defines a collection of *terms* (objects) and *formulas* described in a *Knowledge base* $\mathcal{K}$. For instance, to express the friendship between two terms defined as *Alice* and *Bob*, we can use the predicate `friend_of`:

$$\phi_1 = \texttt{friend\_of}(Alice, Bob) \wedge \texttt{friend\_of}(Bob, Alice)$$

At the same time, we can specify formulas defining general properties, such as the symmetric nature of the friendship relationship within a specific *domain*:

$$\phi_2 = \forall x, y \left(\texttt{friend\_of}(x, y) \Rightarrow \texttt{friend\_of}(y, x)\right)$$

Adopting Real Logic, both formulas and terms are *grounded* (interpreted) into a scalar value in the [0,1] interval. Specifying the grounding function $\mathcal{G}$, which maps terms and formulas into such real-valued features, generates a complete definition of a theory. Given a set of terms, aggregate formulas can be defined by approximating unary, binary or quantifiers connectives in fuzzy logic using suitable differential operators.

In semantic image interpretation tasks, terms (objects) are typically grounded by features computed by a pre-trained convolutional neural network; it is also possible to jointly train the convolutional backbone and the LTNs in an end-to-end fashion [4]. Predicates symbols $p \in \mathcal{P}$ are grounded by a function $\mathcal{G}\left(D(p)\right) \rightarrow [0, 1]$. A typical predicate in semantic image interpretation is the `isOfClass` one, which represents the probability that a given object belongs to class c.

In conventional LTNs [5], [13], [4], predicates are typically defined as the generalization of the neural tensor network:

$$\mathcal{G}\left(\mathcal{P}\right)\left(\mathbf{v}\right) = \sigma\left(u_P^T \tanh\left(\mathbf{v_T} W_P^{[1:k]} \mathbf{v} + V_P \mathbf{v} + b_p\right)\right) \quad (1)$$

where σ is the sigmoid function, $W[1:k] \in \mathbb{R}^{k \times mn \times mn}$, $V_p \in \mathbb{R}^{k \times mn}$, $u_p \in \mathbb{R}^k$, and $b_p \in \mathbb{R}$ are learnable tensors of parameters. For multi-class problems, the sigmoid function could be substituted by a softmax layer to enforce mutual exclusivity [5].

This grounding requires to add an additional predicate for each class (e.g., `isDog`, `isPerson`, etc.), which is embedded into a tensor network with separate weights. Additionally, since class symbols are not grounded, predicates can only be defined for object instances, which rapidly leads to very large knowledge bases when background logical axioms need to be imposed. On the contrary, our proposed grounding does not require additional model parameters, or in any case limits them to a small set which is shared among all `isOfClass` predicates. Furthermore, it encodes abstract classes as parametric objects that live in the same embedding space as their instances, and can be used to establish relationships with other objects (e.g., macro-category relationships). This formulation thus supports more efficient and compact representations.

The *best satisfability* problem, which is the optimization problem underlying LTNs, consists in determining the values of Θ^* that maximize the truth values of the conjunction of all formulas $\phi \in \mathcal{K}$:

$$\Theta^* = argmax_\Theta \hat{\mathcal{G}}_\theta \left(\bigwedge_{\phi \in \mathcal{K}} \phi\right) - \lambda ||\Theta||_2^2 \quad (2)$$

where $\lambda ||\Theta||_2^2$ is a convenient regularization term.

C. Zero-shot learning

In zero-shot learning, a learner must be able to recognize objects from test classes, not seen during training, by leveraging some sort of description, most commonly a vector of semantic attributes [18]. In this paper, we target the Generalized zero-shot learning (GZSL) scenario, in which both seen and unseen classes appear at test time [18]. State-of-the-art techniques for

ZSL classification typically fall within two categories [18], [8]: *embedding-based* and *generative-based*.

Embedding-based models [8], [19], [20], [21] compare semantic characteristics (e.g., attributes) and visual characteristics (usually taken from a pre-trained convolutional neural network) by (learning a) mapping to a common embedding space. Mapping the semantic space to the more compact visual feature space, rather than the opposite, alleviates the so-called hubness problem and facilitates separation between classes [8]. Standard embedding-based models are completely agnostic to any information about the test set: neither examples (even unlabelled), nor class attributes are assumed to be available at training time. Although based on a NeSy formulation, the proposed PROTO-LTN approach can be regarded as an embedding-based technique, as semantic concepts and visual features are mapped onto a common embedding space.

Embedding-based models tend to be naturally biased towards seen classes. To alleviate this problem, generative models were proposed with the purpose of learning a conditioned probability distribution for each class, and thus generate artificial examples of unseen classes [22], [23], [24]. A conventional classifier is trained by utilizing both the true and the generated examples. Although impressive results, especially in a GZSL context, can be achieved by taking advantage of this machinery, reduced flexibility with respect to embedding methods is entailed, as unseen classes need to be defined, so that a number of corresponding examples can be artificially synthesized. PROTO-LTNs are thus best compared with other embedding-based models, although nothing prevents them from being trained on, or combined with, generative methods.

III. PROTOTYPICAL LOGIC TENSOR NETWORKS

First, we introduce the basic notations related to prototypical networks in the FSL (Section III-A) and ZSL (Section III-B) settings [7]. Then, in Sections III-C and III-D, we build on these concepts and show how the PROTO-LTN training cycle is constructed by substituting the original model with a grounded $\mathcal{K}$, and the original loss with a best satisfiability problem.

A. Prototypical Networks: the FSL setting

A N-way-K-shot FSL scenario is supposed, in which a classifier is asked to discriminate the right class among N choices, while having the chance to observe K examples per class [25], [26], [27]. More specifically, the labelled examples are referred to as the *support* examples, whereas the unlabeled ones as the *query* examples.

The underlying assumption that it exists an embedding space in which elements of different classes are well-scattered, and that it can be mathematically translated into an embedding function f_θ whose parameter θ must be inferred, acting as a mapping

$$f_\theta : \mathbb{R}^D \to \mathbb{R}^M. \tag{3}$$

In Eq. 3, D and M are, respectively, the dimensions of the input space and of the embedding space. Thus, for an example x, $f_\theta(x)$ is the corresponding embedding.

In FSL, a *prototype* for class n is obtained as the mean embedding of the K support examples of class n at train time:

$$p_n = \frac{1}{K} \sum_{\substack{(x^{\tilde{S}}, y^{\tilde{S}}) \in \tilde{S} \\ \text{s.t. } y^{\tilde{S}} = n}} f_\theta(x^{\tilde{S}}). \tag{4}$$

Class prototypes thus need to live in the embedding space, as they embody average features shared by elements of the class they represent. At *training time*, θ is optimized so that the distance between each prototype and the elements of its class is minimized, while the distance between different prototypes is maximized. Finally, classification at *testing time* is performed by assigning each query sample to its nearest prototype.

At testing time, a support set is at disposal of N_S labeled examples $S = \{(x_1^S, y_1^S), ..., (x_{N_S}^S, y_{N_S}^S)\}$, where each $x_i^S \in \mathbb{R}^D$ is the feature vector of an example, and $y_i^S \in C \subset \mathbb{N}$ is the corresponding label. Assuming a N-way-K-shot scenario, exactly K support examples are available for each of the N classes. A query set $Q = \{x_1^Q, ..., x_{N_Q}^Q\}$ of N_Q unlabeled examples is thus supplied, and the task is to correctly assort the examples into their classes. The elements from the query set Q belong to the same domain as those from the support set S.

At training time, it could be impossible to know which classes will the testing scenario yield. In other words, a support set S is not accessible in advance. To cope with that, a training set $T = \{(x_1^T, y_1^T), ..., (x_{N_T}^T, y_{N_T}^T)\}$ is chosen that reflects the best prior information possessed about the testing scenario, with labels $y_i^T \in C_T \subset \mathbb{N}$ and $|C_T| = N_T$ classes which can coincide or outnumber them ($N_T \geq N$). In other words, it is possible that $C \cap C_T \neq \emptyset$, but it cannot be said in advance. Then, *fake* support and query sets $\tilde{S} \subset T$ and $\tilde{Q} \subset T$ are extracted to mimic the testing scenario and instruct the model to learn accordingly.

B. Prototypical networks: the ZSL setting

In ZSL, one does not dispose of labelled examples for all classes. Instead, it is assumed that N abstract vectors denoted as $\{a^{(1)}, a^{(2)}, ..., a^{(N)}\}$, with $a^{(n)} \in \mathbb{R}^A$, encode the characteristics of all N classes.

As in FSL, at training time one takes advantage of a set $T = \{(x_1^T, y_1^T), ..., (x_{N_T}^T, y_{N_T}^T)\}$ of labelled examples from classes $y_i^T \in C_T \subset \mathbb{N}$, where it is preferably $|C_T| = N_T \geq N = |C|$. The training cycle remains unchanged in the ZSL case, but class prototypes are defined differently:

- the embedding for a query example x^Q is still obtained as $f_\theta(x^Q)$, where $f_\theta : \mathbb{R}^D \to \mathbb{R}^M$;
- the prototype for class $n \in C$ is extracted as $p_n = g_\theta(a^{(n)})$ via a separate embedding function $g_\theta : \mathbb{R}^A \to \mathbb{R}^M$, which maps the semantic attribute space to the common embedding space.

C. PROTO-LTN: the FSL scenario

The overall architecture of PROTO-LTN, when tailored to the ZSL scenario, is illustrated in Fig. 1. The input image embeddings are extracted from a convolutional neural network, while attribute vectors are mapped into the embedding domain through an embedding function. In this section, details about the definition of the grounding of the constant, variables, functions and predicates are given. Then, the Knowledge Base $\mathcal{K}$ which encodes our learning problem is defined.

1) Groundings terms: Within a single training episode, a batch of training samples is selected in the form of fake support $\tilde{S}$ and query $\tilde{Q}$ sets. Groundings for variables and their domain D (not learnable) can be defined as

$$\begin{aligned}
\mathcal{G}(q) &= \langle x_1^{\tilde{Q}}, ..., x_{N_{\tilde{Q}}}^{\tilde{Q}} \rangle, && (5)\\
\mathcal{G}(q_l) &= \langle y_1^{\tilde{Q}}, ..., y_{N_{\tilde{Q}}}^{\tilde{Q}} \rangle, && (6)\\
\mathcal{G}(q_e) &= \mathcal{G}(\texttt{getEmbedding}(q)) && (7)\\
&= \langle f_\theta(x_1^{\tilde{Q}}), ..., f_\theta(x_{N_{\tilde{Q}}}^{\tilde{Q}}) \rangle, && (8)\\
\mathcal{G}(s) &= \langle x_1^{\tilde{S}}, ..., x_{N_S}^{\tilde{S}} \rangle, && (9)\\
\mathcal{G}(s_l) &= \langle y_1^{\tilde{S}}, ..., y_{N_S}^{\tilde{S}} \rangle, && (10)\\
\mathcal{G}(p),\, \mathcal{G}(p_l) &= \mathcal{G}(\texttt{getPrototypes}(s, s_l)) && (11)\\
&= \Pi_\theta(\mathcal{G}(s, s_l)) && (12)\\
&= \Pi_\theta(\langle (x_1^{\tilde{S}}, y_1^{\tilde{S}}), ..., (x_{N_S}^{\tilde{S}}, y_{N_S}^{\tilde{S}}) \rangle), && (13)
\end{aligned}$$

where q are the query examples ($D(q) = \texttt{features}$), q_l are the corresponding labels ($D(q_l) = \texttt{labels}$), and q_e are their embeddings ($D(q_e) = \texttt{embeddings}$). Conversely, s are the examples in the support set ($D(s) = \texttt{features}$) and s_l their labels. Finally, p and p_l are the prototypes and their labels, respectively, with $D(p) = \texttt{embeddings}$ and $D(p_l) = \texttt{labels}$.

2) Grounding functions and predicates: PROTO-LTNs are based on two functions (`getEmbedding` and `getPrototypes`) and the `isOfClass` predicate.

`getEmbedding` is a conventional LTN function which maps image features into the embedding space, hence $D_{\text{in}}(\texttt{getEmbedding}) = \texttt{features}$ to $D_{\text{out}}(\texttt{getEmbedding}) = \texttt{embeddings}$.

The `getPrototypes` function, with $D_{\text{in}}(\texttt{getPrototypes}) = \texttt{features} \times \texttt{labels}$ and $D_{\text{out}}(\texttt{getPrototypes}) = \texttt{embeddings} \times \texttt{labels}$, returns labelled prototypes given a support set of labelled examples. Each prototype is in fact a function of all support points belonging to the same class, as defined in Eq. 4. It is defined as a *generalized* LTN function, which accepts as input multiple instantiations of variables (and hence multiple domains).

Groundings for both functions are defined as:

$$\mathcal{G}(\texttt{getEmbedding}) = f_\theta, \tag{14}$$

$$\mathcal{G}(\texttt{getPrototypes}) = \Pi_\theta, \tag{15}$$

where $f_\theta : \mathbb{R}^D \to \mathbb{R}^M$ defines the embedding function, whereas

$$\Pi_\theta : \bigcup_{l=1}^{\infty} {}_{m=1}^{l} \mathbb{R}^D \times \mathbb{N} \to \bigcup_{l=1}^{\infty} {}_{m=1}^{l} \mathbb{R}^M \times \mathbb{N}$$

accepts as input a list of N_S labelled support examples, i.e., an element of $(\mathbb{R}^D \times \mathbb{N})^{N_S}$, and returns a list of labelled prototypes for all the $\tilde{N}$ classes seen in the support set, or an element of $(\mathbb{R}^M \times \mathbb{N})^{\tilde{N}}$.

The `isOfClass` predicate for class $n \in C$ is grounded as:

$$\mathcal{G}(\texttt{isOfClass}) = e^{-\alpha\, d(\cdot,\cdot)^2}, \tag{16}$$

where α is a hyperparameter and d is a measure of distance. $\mathcal{G}(\texttt{isOfClass}) : \mathbb{R}^M \times \mathbb{R}^M \to [0, 1]$; $\mathcal{G}(\texttt{isOfClass})$takes the value of 1 when the distance from the class prototype $d(\cdot, \cdot)$ is 0. In our formulation the Euclidean distance squared is adopted, as in DEM [8]. Alternatively, parametric similarity functions could be used:

$$\mathcal{G}''(\texttt{isOfClass}) = \sigma_\theta(\text{Concatenate}[\cdot, \cdot]). \tag{17}$$

where σ_θ could be a MLP with output sigmoid activation. This formulation is closer to that of Relation Networks [19].

3) Knowledge Base: $\mathcal{K}$ represents our knowledge about the formulated problem and is updated at each training episode based on the current fake support set. $\mathcal{K} = \{\phi_{\text{aff}}, \phi_{\text{neg}}\}$ contains two aggregations of formulas which specify that each query item is a positive example for its class, and a negative one for all the others:

$$\begin{aligned}
\phi_{\text{aff}} = \forall \text{Diag}(q_e, q_l)(\forall \text{Diag}(p, p_l)\\
: q_l = p_l(\texttt{isOfClass}(q_e, p))),
\end{aligned} \tag{18}$$

$$\begin{aligned}
\phi_{\text{neg}} = \forall\, \text{Diag}(q_e, q_l)\, (\forall\, \text{Diag}(p, p_l)\\
: q_l \neq p_l\, (\neg \texttt{isOfClass}(q_e, p))).
\end{aligned} \tag{19}$$

We have exploited both Diagonal Quantification and Guarded Quantifiers, whose formal definition can be found in [5].

PROTO-LTN is trained by maximizing the satisfiability

$$\mathcal{L}^{\text{ep}} = 1 - \left(\bigwedge_{\phi \in \mathcal{K}} \phi \right) = -\mathcal{G}(\phi_{\text{aff}}) - w_{\text{n}}\, \mathcal{G}(\phi_{\text{neg}}), \tag{20}$$

where the weight w_{n} reflects the expectation that negations play a less discriminative role than affirmation in classification. In our experiments, we set $w_{\text{n}} = 0$ and consider only ϕ_{aff}, leaving exploration of this hyper-parameter to future work.

By introducing an aggregation function [5], [11], we obtain

$$\mathcal{L}^{\text{ep}} = \left(-\log(\mathcal{G}(\phi_{\text{aff}}))^{\frac{1}{p_{\text{agg}}}}) + w_{\text{n}}(1 - \mathcal{G}(\phi_{\text{n}}))^{\frac{1}{p_{\text{agg}}}} \right)^{p_{\text{agg}}} \tag{21}$$

where $\mathcal{G}(\phi_{\text{aff}})$ is implemented through the generalized product p-mean operator and $\mathcal{G}(\phi_{\text{neg}})$ with the generalized mean operator A_{pM}:

$$A_{pPR}(\tau_1, ..., \tau_n) = \left(\prod_{i=1}^{n} \tau_i \right)^{\frac{1}{p_\forall}}, A_{pM}(\tau_1, ..., \tau_n) = \left(\frac{1}{n} \sum_{i=1}^{n} \tau_i^p \right)^{\frac{1}{p_\forall}}.$$

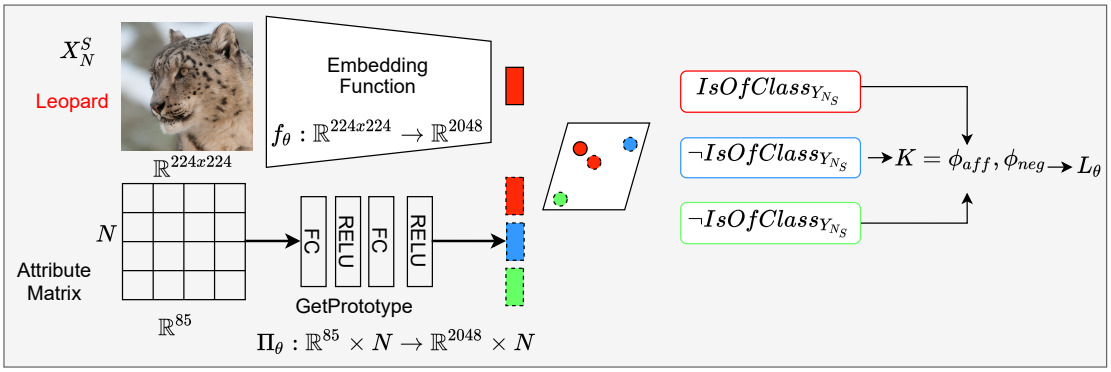

Fig. 1. Proto-LTN architecture for ZSL classification. The architecture is composed of a convolutional features extractor and an attribute encoder. The two branches allow to map semantic and visual features in a common embedding space. The `isOfClass` predicate aims to minimize the distance between instances (solid line circles) and class prototypes (dashed line circles) based on affirmative and negative formulas embedded in the knowledge base $\mathcal{K}$. At train time, the loss function maximizes the satisfiability (truth value) of all formulas in $\mathcal{K}$.

It should be noticed that the choice of p_{agg} does not need to coincide with that of $p_\forall$ for quantification, and both hyper-parameters need to be tuned experimentally.

When optimizing a positive quantity, a common practice consists in optimizing its logarithm: the product between similarities takes a more desirable form when A_{pPR} is used as the aggregation operator for $\forall$. Unfortunately, one does not obtain an equally appealing expression for ϕ_{neg}.

If a squared Euclidean distance is used as similarity measure and the negation weight w_{n} is set to 0, one obtains the same formulation of the loss function of DEM [8], up to a scaling constant:

$$\mathcal{L}^{\text{ep}} = -\log\left(e^{-\frac{\alpha}{p_\forall}\left(\sum_{n\in\tilde{C}}\sum_{\substack{(x^{\tilde{Q}},y^{\tilde{Q}})\in\tilde{Q}\\ \text{s.t. } y^{\tilde{Q}}\neq n}} d(f_\theta(x^{\tilde{Q}}),p_n)^2\right)}\right)$$
$$= \frac{\alpha}{p_\forall}\left(\sum_{n\in\tilde{C}}\sum_{\substack{(x^{\tilde{Q}},y^{\tilde{Q}})\in\tilde{Q}\\ \text{s.t. } y^{\tilde{Q}}\neq n}} d(f_\theta(x^{\tilde{Q}}),p_n)^2\right). \quad (22)$$

D. PROTO-LTN: the GZSL scenario

The GZSL setting is analogous to the FSL setting, with the main difference lying in how prototypes are defined and calculated. No generalized LTN functions are needed for the GZSL case. Computations for a training epoch are reported in Algorithm 1.

Since only one semantic vector $a^{(n)}$ is given for each class n, there is a 1-to-1 correspondence between elements of the support set and prototypes. The latter are embodied by the semantic embedding function $g_\theta : \mathbb{R}^A \rightarrow \mathbb{R}^D$ obtaining as the feature space the common embedding space. We just define `getPrototypes` as a conventional LTN function, whose grounding is $\mathcal{G}(\texttt{getPrototypes}) = g_\theta$. Conversely, nothing changes for the query map `getEmbedding`.

Algorithm 1 PROTO-LTN - GZSL Training procedure

function TRAIN
  Input $\leftarrow q$ Training Images
  Input $\leftarrow q_l$ Training label
  Input $\leftarrow a$ Semantic attribute set
  Input $\leftarrow a_l$ Semantic attribute label
  for i *in* $N_{TrainingSteps}$ **do**
    $q_{ei} \leftarrow \texttt{getEmbedding}(q)$
    $a_i, a_{li} \leftarrow \texttt{getAttributes}(a)$
    $p_i, p_{li} \leftarrow \texttt{getPrototypes}(a_i, a_{li})$
    $\phi_{\text{aff}} = \forall\text{Diag}(q_{ei}, q_{li})(\forall\text{Diag}(p_i, p_{li}) : q_{li} = p_{li}(\texttt{isOfClass}(q_{ei}, p_i)$
    $\phi_{\text{n}} = \forall\,\text{Diag}(q_i, q_{li})\,(\forall\,\text{Diag}(p_i, p_{li}) : q_{li} \neq p_{li}\,(\neg\texttt{isOfClass}(q_{ei}, p_i))$
    $\mathcal{L}^{\text{ep}} = \left(-\log((\mathcal{G}(\phi_{\text{aff}}))^{\frac{1}{p_{\text{agg}}}}) + w_{\text{n}}(1-\mathcal{G}(\phi_{\text{n}}))^{\frac{1}{p_{\text{agg}}}}\right)^{p_{\text{agg}}}$
    $computeGradient(\mathcal{L}^{\text{ep}})$
    $updateGradient$
  end for
function TEST
  Input $\leftarrow q$ Test Images
  Input $\leftarrow a$ Semantic attribute set
  $q_e \leftarrow \texttt{getEmbedding}(q)$
  $a, a_l \leftarrow \texttt{getAttributes}(a)$
  $p, p_l \leftarrow \texttt{getPrototypes}(a, a_l)$
  for i *in* len(q_e) **do**
    for j *in* len(p) **do**
      $prediction_i \leftarrow \texttt{isOfClass}(q_{ei}, p_j)$
    end for
  end for

IV. Experimental Settings

Experiments were conducted in both ZSL and GZSL settings on the Awa2 (Animals with Attributes) [18], CUB [28], aPY (Attribute Pascal and Yahoo)[29] and SUN (Scene Understanding) [30] benchmarks. For all datasets, image encodings, attributes and splits were collected from the original benchmark [18].

The entire architecture is composed of two different blocks: the image visual encoder and the semantic encoder. The embedding function f_θ is composed by a ResNet101 [33] embedding model, pretrained on ImageNet [34] and kept frozen, which converts an image I into a vector $\mathbf{x} \in \mathbb{R}^M$, where $M = 2048$. This setting is maintained in all experiments

TABLE I

FOR PROTO-LTN, WE SHOW MEAN $\pm$ STANDARD DEVIATION AND MAXIMUM (IN PARENTHESIS) PERFORMANCE. TOP1^{ZSL} (T1), $\text{TOP1}^{\text{GZSL\_UNSEEN}}$ (U), $\text{TOP1}^{\text{GZSL\_SEEN}}$ (S) AND H^{GZSL} (H) ARE ALWAYS OBTAINED ON THE PROPOSED SPLIT (PS) OF AWA2, CUB, APY AND SUN CLASSES, AS DESCRIBED IN [18]. † ASSUMES A TRANSDUCTIVE ZSL SETTING. BEST PERFORMANCES ARE REPORTED IN BOLD.

Method	Awa2				CUB				APY				SUN			
	T1	U	S	H	T1	U	S	H	T1	U	S	H	T1	U	S	H
SYNC (2016) [31]	46.6	10.0	90.5	18.0	55.6	11.5	**70.9**	19.8	-	-	-	-	56.3	7.9	43.3	13.4
Relation Net (2017)[19]	64.2	30.0	93.4	45.3	55.6	38.1	61.1	47	-	-	-	-	-	-	-	-
PrEN† (2019) [32]	74.1	32.4	88.6	47.4	66.4	35.2	55.8	43.1	-	-	-	-	**62.9**	**35.4**	27.2	**30.8**
VSE (2019) [20]	**84.4**	**45.6**	**88.7**	**60.2**	**71.9**	**39.5**	68.9	**50.2**	**65.4**	**43.6**	**78.7**	**56.2**	-	-	-	-
DEM (2017) [8]	67.1	30.5	86.4	45.1	51.7	19.6	57.9	29.2	35.0	11.1	75.1	19.4	61.9	20.5	34.3	25.6
PROTO-LTN	67.6	32.0	83.7	46.2	48.8	20.8	54.3	30.0	35.0	17.1	66.2	27.21	60.4	20.4	**36.8**	26.2
	±1.1	±1.3	±0.3	±1.3	±1.2	±2.6	±1.1	±3.0	±3.1	±2.0	±5.1	±2.9	±2.5	± 1.0	±4.4	±1.9
	(70.8)	(34.8)	(84.3)	(49.1)	(50.3)	(23.4)	(55.7)	(33.0)	(38.6)	(19.4)	(70.7)	(30.0)	(62.1)	(22.15)	(39.9)	(28.0)

with all datasets.

Semantic vectors are encoded in the embedding space via a function g_θ, which consists of two fully connected layers (FC) with ReLU activation function, initialized by a truncated normal distribution function. We set the hyper-parameter aggregations to $p_{agg} = 1$ and $p_\forall = 2$, also taking into account preliminary experiments on Awa2 [18].

The framework was implemented in Tensorflow based on the LTN package [5], [35]. Experiments were conducted on a workstation equipped with an Intel® Core™ i7-10700K CPU and a RTX2080 TI GPU. All networks were trained for 30 epochs with Adam optimizer and batch size 64. Hyper-parameters (learning rate, α and regularization term λ) were optimized separately for each dataset. Standard performance metrics for GZSL were used as defined in [18]. Mean and standard deviation were calculated by repeating each experiment three times.

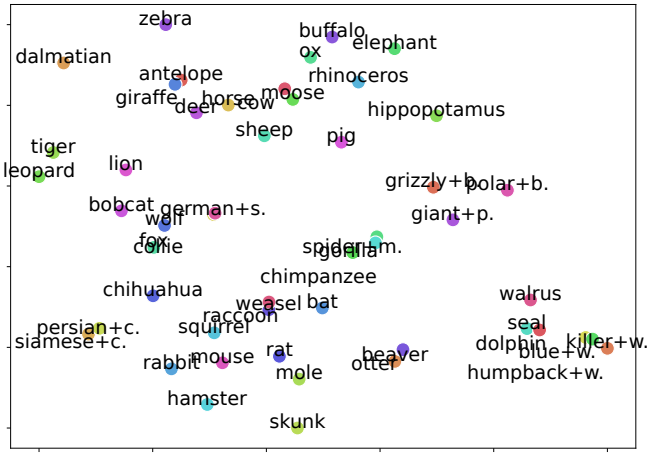

Fig. 2. t-SNE visualization of class prototypes for the Awa2 dataset.

V. RESULTS

PROTO-LTN results are reported in Table I, along with those for comparable embedding-based methods. Fig. 2 illustrates the embedding space with highlighted class prototypes.

As expected based on our analytical analysis, experimental performance is competitive with respect to most embedding-based techniques, in particular DEM [8] and Relation Net [19], which rely on similar assumptions and the same input as the current PROTO-LTN implementation. As shown in Section III-C, under certain conditions the PROTO-LTN loss is equivalent to that of DEM, up to a scaling constant, albeit with different regularization terms. We outperform DEM on unseen classes for all experimental benchmarks: this entails that the proposed formulation is a strong basis for a novel, NeSy approach to the GZSL task.

Our method is outperformed by VSE, which relies on a different strategy to compute visual feature embeddings. A semantic loss allows to align the embedding space with part-feature concepts provided by a semantic oracle. Since the latter relies on an external knowledge base, it contains concepts beyond the available semantic vector $\{a^{(1)}, a^{(2)}, ..., a^{(N)}\}$. This is especially advantageous in benchmarks like aPY, in which attributes are noisy and not visually informative [20]. This is a limitation of our current experiments, but not intrinsic to PROTO-LTNs. Indeed, $\mathcal{K}$ can be extended to include part-of relationships between concepts, and previous works have shown how these relationships can be leveraged to impose symbolic priors during learning, e.g., in object detection [4], [13]. However, the LTN formalism needs to be further extended to align part-based concepts with their visual groundings in an unsupervised fashion.

VI. CONCLUSIONS AND FUTURE WORKS

We introduced PROTO-LTN, a novel Neuro-Symbolic architecture which extends the classical formulation of LTN borrowing from embeddings-based techniques. Following the strategy of PNs, we entirely focus on learning embedding functions (such as f_θ and g_θ), implying that class prototypes are obtained ex-post, based on a support set. These methods are robust to noise, an essential property in FSL, and provide a scheme to embed both examples (images) and class prototypes in the same metric space. This is a key property in the context of LTNs, because it enables different levels of abstraction: one can either state something about a particular example, or about an entire class, as prototypes can be viewed as parametrized labels for classes. We have shown the viability of our approach

in GZSL and leave to future work the extension to other settings (e,g., few-shot or semi-supervised learning).

While our experimental results are encouraging, we argue that the strength of our formulation lies in its generality, and the full potential of PROTO-LTN is yet to be realized. Future work can aim at two complementary directions. First, alternative formulations of the `isOfClass` relationship could be explored, by changing the distance metric and/or the prototype encoding. Mapping class prototypes back to the input space, as done for instance in [36], could improve explainability.

Second, the knowledge $\mathcal{K}$ could be extended to leverage prior information, e.g., from external knowledge bases, to improve generalization to unseen classes. Experiments should include both inductive and transductive settings: the assumption that information about attributes and relationships of unseen classes is available at training or test time (e.g., from WordNet) is less restrictive than assuming that actual examples, albeit unlabelled, are available.

References

[1] L. De Raedt, S. Dumancic, R. Manhaeve, and G. Marra, "From statistical relational to neuro-symbolic artificial intelligence," in *29th International Joint Conference on Artificial Intelligence*, 2021, pp. 4943–4950.

[2] T. R. Besold, A. d. Garcez, S. Bader, H. Bowman, P. Domingos, P. Hitzler, K.-U. Kühnberger, L. C. Lamb, D. Lowd, P. M. V. Lima *et al.*, "Neural-symbolic learning and reasoning: A survey and interpretation," *arXiv preprint arXiv:1711.03902*, 2017.

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

[4] F. Manigrasso, F. D. Miro, L. Morra, and F. Lamberti, "Faster-LTN: a neuro-symbolic, end-to-end object detection architecture," in *International Conference on Artificial Neural Networks*. Springer, 2021, pp. 40–52.

[5] S. Badreddine, A. d. Garcez, L. Serafini, and M. Spranger, "Logic tensor networks," *Artificial Intelligence*, vol. 303, p. 103649, 2022.

[6] L. Serafini, A. d'Avila Garcez, S. Badreddine, I. Donadello, M. Spranger, and F. Bianchi, "Logic tensor networks: Theory and applications," in *Neuro-Symbolic Artificial Intelligence: The State of the Art*. IOS Press, 2021, pp. 370–394.

[7] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *31st International Conference on Neural Information Processing Systems*, 2017, pp. 4080–4090.

[8] L. Zhang, T. Xiang, and S. Gong, "Learning a deep embedding model for zero-shot learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3010–3019.

[9] Z. Wan, D. Chen, Y. Li, X. Yan, J. Zhang, Y. Yu, and J. Liao, "Transductive zero-shot learning with visual structure constraint," *Advances in Neural Information Processing Systems*, vol. 32, pp. 9972–9982, 2019.

[10] A. Goyal and Y. Bengio, "Inductive biases for deep learning of higher-level cognition," *arXiv preprint arXiv:2011.15091*, 2020.

[11] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of Machine Learning Research*, vol. 9, no. 11, 2008.

[12] D. Yu, B. Yang, D. Liu, and H. Wang, "A survey on neural-symbolic systems," *arXiv preprint arXiv:2111.08164*, 2021.

[13] I. Donadello, L. Serafini, and A. D. Garcez, "Logic tensor networks for semantic image interpretation," in *26th International Joint Conference on Artificial Intelligence*, 2017, pp. 1596–1602.

[14] K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, and J. B. Tenenbaum, "Neural-symbolic VQA: disentangling reasoning from vision and language understanding," in *32nd International Conference on Neural Information Processing Systems*, 2018, pp. 1039–1050.

[15] R. Vedantam, K. Desai, S. Lee, M. Rohrbach, D. Batra, and D. Parikh, "Probabilistic neural symbolic models for interpretable visual question answering," in *International Conference on Machine Learning*, 2019, pp. 6428–6437.

[16] Z. Li, E. Stengel-Eskin, Y. Zhang, C. Xie, Q. H. Tran, B. Van Durme, and A. Yuille, "Calibrating concepts and operations: Towards symbolic reasoning on real images," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 910–14 919.

[17] M. van Bekkum, M. de Boer, F. van Harmelen, A. Meyer-Vitali, and A. ten Teije, "Modular design patterns for hybrid learning and reasoning systems," *Applied Intelligence*, pp. 1–19, 2021.

[18] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning — A comprehensive evaluation of the good, the bad and the ugly," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2251–2265, 2019.

[19] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1199–1208.

[20] P. Zhu, H. Wang, and V. Saligrama, "Generalized zero-shot recognition based on visually semantic embedding," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[21] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "DeViSE: A deep visual-semantic embedding model," in *26th International Conference on Neural Information Processing Systems*, 2013, p. 2121–2129.

[22] V. K. Verma, G. Arora, A. Mishra, and P. Rai, "Generalized zero-shot learning via synthesized examples," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[23] H. Huang, C. Wang, P. S. Yu, and C.-D. Wang, "Generative dual adversarial network for generalized zero-shot learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[24] Y. Xing, S. Huang, L. Huangfu, F. Chen, and Y. Ge, "Robust bidirectional generative network for generalized zero-shot learning," in *IEEE International Conference on Multimedia and Expo*, 2020, pp. 1–6.

[25] E. G. Miller, N. E. Matsakis, and P. A. Viola, "Learning from one example through shared densities on transforms," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2000, pp. 464–471.

[26] B. Lake, R. Salakhutdinov, J. Gross, and J. Tenenbaum, "One shot learning of simple visual concepts," *Cognitive Science*, vol. 33, 2011.

[27] G. Koch, R. Zemel, R. Salakhutdinov *et al.*, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, vol. 2. Lille, 2015.

[28] C. Wah, S. Branson, P. Perona, and S. J. Belongie, "Multiclass recognition and part localization with humans in the loop," *International Conference on Computer Vision*, pp. 2524–2531, 2011.

[29] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1778–1785.

[30] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "SUN database: Large-scale scene recognition from abbey to zoo," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3485–3492.

[31] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha, "Synthesized classifiers for zero-shot learning," in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5327–5336.

[32] M. Ye and Y. Guo, "Progressive ensemble networks for zero-shot recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[34] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[35] S. Badreddine, A. Garcez, L. Serafini, and M. Spranger, "GTS: Logic Tensor Network library," https://github.com/logictensornetworks/logictensornetworks, 2021.

[36] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, "This looks like that: Deep learning for interpretable image recognition," *Advances in Neural Information Processing Systems*, vol. 32, pp. 8930–8941, 2019.