Visual Question Answering with DeepProbLog

Jorrit Willaert[1]
[1]Catholic University of Leuven, Leuven – Belgium
e-mail: jorrit.willaert@student.kuleuven.be

***Abstract* – TODO**

***Keywords* – Neuro Symbolic AI, Visual Question Answering, DeepProbLog, Problog, Convolutional Neural Networks**

I. INTRODUCTION

The Neuro Symbolic AI community is growing rapidly, indicating that people start to recognize the value of the field. The Neuro Symbolic AI field is interested in building a bridge between the robustness of probabilistic knowledge, with the well-known popularity and proven strengths of deep neural networks. DeepProbLog [1] offers this ability, by using both the strengths of neural networks (i.e. system 1, typical subconscious tasks such as visual recognition, the processing of languages, ...), along with the strengths of rule-based probabilistic systems (i.e. system 2, slow, sequential thinking such as the derivation of a proof).

This paper elaborates on an application that requires both systems to be used, namely Visual Question Answering. System 1 will be required in order to gain an understanding of the image under investigation, with in particular their shapes and colors. System 2, on the other hand, will use this extracted information for deriving certain properties of objects[1], or even for capturing the relations[2] between the objects.

II. LITERATURE SURVEY

The application focuses on Visual Question Answering (VQA), for which huge datasets are present, along with very sophisticated methods. The best known dataset for VQA is CLEVR [2], which contains 100k images with one million questions. An example image is given in Figure 1, while example questions are:

- Are there an equal number of large things and metal spheres?
- What size is the cylinder that is left of the brown metal thing that is left of the big sphere?
- How many objects are either small cylinders or metal things?

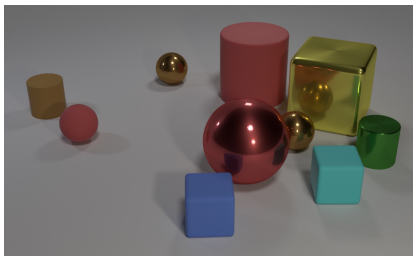

Fig. 1. A sample image from the CLEVR dataset [2]

Clearly, both system 1 and system 2 are actively used when answering these questions. One could wonder if neural networks alone could answer these questions without having an explicit system 2 encoding (i.e. the rule based knowledge base). Intuitively, it makes sense that if certain facts of the world are known[3], learning can proceed much more quickly[4]. Seen from an optimization viewpoint, errors made during prediction in this setup can be targeted exactly, which makes the optimization process also more targeted, and hence more efficient. Finally, this paper also provides evidence for these statements, since in Section Athe comparison between a VQA implementation with DeepProbLog is made with a purely CNN based approach.

This paper is based on the CLEVR dataset, but uses however a much more simplified version. In essence, it is almost like the Sort-Of-CLEVR dataset [3]. This Sort-Of-CLEVR dataset contains images such as in Figure 2, while asking questions such as:

- Non-relational questions: the shape, horizontal or vertical location of an object.
- Relational questions: shape of the closest/furthest object to the object under investigation, or the number of objects with the same shape.

[1]For example, finding the shape of the green object, or deriving if it is located on the left hand side of the image.

[2]Here, one could think of deriving if an object is located to the left of another object, or also for finding the number of circles in the image.

[3]They can be encoded, e.g. counting the number of spheres is simply a matter of detecting all the spheres in the image, after which a mathematical summation is a statement in the knowledge base.

[4]Not to say that learning might even be impossible if a lot of background knowledge is required.

Fig. 2. A sample image from the CLEVR dataset [3]

As explained earlier, both system 1 and system 2 are required for these types of VQA's.

Finally, since this application uses DeepProbLog, quite some time was spent in digesting the DeepProbLog paper [1], along with understanding the examples provided in the code repository [4].

III. APPROACH

The implementation process involved three main parts:

1. Generation of data.
2. Creation of the logical part with DeepProbLog statements.
3. Linking the data and controlling the training process in pure Python code.

A. Generation of data

As mentioned in Section II, the data used in this application is based on the Sort Of CLEVR dataset, with one extra simplification. Given that the logical part will have to decide whether an object is located on the left side of an image, the neural network will have to convey positional information to the logical part. Hence, each discrete position will have to be encoded by a possible outcome of the neural network. Therefore, objects may only be located at certain places in a grid. In this application, a grid of 2x2, 4x4 and 6x6 has been used.

The data generator that was used for the creation of the Sort Of CLEVR dataset has been modified in order to place objects in the mentioned grid positions [3]. An example of a generated image is given in Figure 3.

Fig. 3. A sample image from the dataset that has been used for this application

Each specified color will have an object located somewhere in the grid, of which the shape can be a square or a circle.

These images are accompanied with three non-binary questions and one binary question. For each question, a random object is chosen. The possibilities are:

1. Non-binary - What is the shape of this object [5]?
2. Non-binary - Is this object located on the left side of the image?
3. Non-binary - Is this object located on the bottom side of the image?
4. Binary - How many objects have the same shape of this object?

These questions are encoded in a one-hot encoding, after which they are stored in a CSV file, along with the expected answers. A training set, a validation set, and a test set are generated.

B. Controlling the training process

The overall training process is controlled via the Python interfaces of DeepProbLog, along with general PyTorch implementations of the CNN's. First of all, CNN's are defined with PyTorch. A relatively simple network is used, where the input is given as a square RGB image of 100 pixels wide, which is transformed by the CNN into 72 output features for the 6x6 grid. Each color that is present in the image has its accompanied CNN network, hence the 72 output features encode the possible positions of the object with that color, along with their shape, which can be either square or circular ($6 \cdot 6 \cdot 2 = 72$).

The final thing (besides the logical rule encodings) required before commencing the training process, are the data loaders. The most challenging part here is the transformation from the generated data to specific query mappings and their outcome. One of the questions are chosen, and a correct mapping between a predicate in the logical rule encoding file and the answer is set up.

C. Logical rule encodings

Once the CNN belonging to a specific color has determined the position and the shape of that object, logical rules can deduce whether this object is located on the left hand side of the image, on the bottom side, and how many objects have the same shape. The logical rule program has been listed in Appendix A.

IV. EXPERIMENTS

A. COMPARISONS WITH PURE SYSTEM 1 APPROACHES

The network based on pure neural predicates is able to recognize the questions quickly, however, seems to experience difficulties when having to decide for the correct answer. This can clearly be seen in the confusion matrix.

V. APPENDIX

A. Logical rule encodings

VI. ORGANIZATION OF THE PAPER

This section presents the main issues for editing the manuscript.

A. General Organization

The papers that shall be published in the Brazilian Power Electronics Journal must contain the following main sections: 1) Title; 2) Authors and Affiliations; 3) Abstract and Keywords; 4) Introduction; 5) Body Text; 6) Conclusions; 7) References; 8) Biographies. This order must be respected,

[5] I.e. the shape will be either a square or a circle.

unless the authors add some items, such as: Nomenclature; Appendices and Acknowledgements.

Some comments regarding the main items of the manuscripts are presented below.

1) Title: The title of the paper should be as succinct as possible, stating the subject of the paper in a very clear manner. It should be centered at the top of the first page, in bold, type size 14 points, with the whole title in capital letters.

2) Authors and affiliations: Below the title (leaving one blank line), also centered, the name(s) of the author(s) must be included. The middle names may be abbreviated, but the first and last names must be written in their complete forms (type size 12 points). Immediately below the authors' names, their affiliations, with city, state and country, must be informed (type size 10 points). The electronic addresses must be informed just below the affiliations (type size 10 points).

3) Abstract and keywords: This part is considered one of the most important in the whole paper. It is based on information in Abstract and Keywords that technical papers are indexed and stored in databases.

The Abstract should have no more than 200 words, indicating the main ideas contained in the paper, as well as procedures and obtained results. The Abstract should not be confused with the Introduction and should not have any abbreviations, references, figures, etc. For writing the Abstract, as well as the whole manuscript, you should use passive voice, e. g., "... the experimental results show that..." instead of "... the results we obtained show that...". The word Abstract must be written both in italic and in bold. The Abstract text should be in bold.

Keywords are index terms that identify the main topics of the paper. The term Keywords must be both in italic and bold. The Keywords themselves should be in bold.

4) Introduction: The Introduction must prepare the reader for the paper he/she will read, including a historical overview of the subject and also presenting the main contributions of the paper. The Introduction must not be similar to the Abstract and it is the first section of the paper to be numbered as a section.

5) Body text: The authors must organize the body text in various sections, which should contain important information about the proposal of the paper, facilitating its comprehension for readers.

6) Conclusions: The conclusions should be as clear as possible, highlighting the importance of the paper in the respective research area. The advantages and disadvantages of the proposed subject should be clearly emphasized, as well as the obtained results and possible applications.

7) References: The citation of references throughout the text should appear between square brackets, just before the punctuation mark at the end of the sentence in which the reference is inserted. Only the number of the references should be used, avoiding citations such as "... according to the reference [2]...".

Papers that were accepted for publication, but were not published yet, should also be in the references along with the citation "in Press".

Papers from journals and conferences must begin with the name of the authors (initials followed by the last name), followed by the title, journal or conference name (in italic), number of volume, pages, month and year of publication.

Regarding books, following the name of the authors (initials followed by last name), the title should be in italic, and then should come the publisher, number of edition and place and year of publication.

At the end of these guidelines, there is an example of how the references should be inserted

8) Biographies: The biographies of the authors should appear in the same order as in the beginning of the paper and should basically contain the following items:

- Full name (in bold and underlined);
- Place and year of undergraduation and graduation conclusions;
- Professional experience (Institutions and companies in which they have worked, number of patents obtained, areas of expertise, relevant scientific activities, scientific societies in which they are members, etc.).

In case additional items are used, such as Nomenclature, Appendices and Acknowledgements, the following instructions should be considered:

9) Nomenclature: The nomenclature consists of the definition of quantities and symbols used throughout the paper. Its inclusion is not mandatory and this item must not be numbered. If this item is included, it should precede the Introduction. In case the authors do not include this item, the definition of quantities and symbols must occur during the text, right after they appear. In the beginning of these guidelines there is an example of this optional item.

10) Acknowledgements and appendices: The acknowledgements to any collaborators, as well as appendices, do not receive any numeration and should be at the end of the text, before the references. At the end of this text there is an example of this optional item.

On the last page of the paper, the authors should distribute the contents evenly, using both columns, in a way that both end in a parallel manner.

B. Organization of the Sections of the Paper

The organization of the manuscript in titles and subtitles is important to divide it in sections, which help the reader to find subjects of interest in the paper. They also help the authors to develop their paper in an orderly form. The paper can be organized in primary, secondary and tertiary sections.

The primary sections are the titles of the actual sections. They are written in capital letters in the center of the column separated by a blank line above and another one below them, and sequential Roman numerals should be used.

The secondary sections are the subtitles of the sections. Just the first letter of each word of the section should be written with a capital letter. It should be located at the left part of the

column being separated by a blank line above from the rest of the text. The designation of the secondary sections is done with letters in uppercase form, followed by a dot. They should be in italic.

The tertiary sections are subdivisions of the secondary sections. Only the first letter of the first word of the section should be a capital letter. The designation of the tertiary sections should be done with Arabic numerals, followed by parentheses. They should be in italic.

VII. OTHER INSTRUCTIONS

A. Editorial Rules

For papers with multiple authors, it is necessary to inform the order of presentation of the authors and filling out the Copyright form at the https://mc04.manuscriptcentral.com/revistaep, authorizing the publication of the paper.

The Brazilian Power Electronics Journal should be considered source of original publication. It reserves its right to make normative, spelling and grammatical modifications in the original files, but respecting the style of the authors. The final versions cannot be sent to the authors.

The published papers will become property of the Brazilian Power Electronics Journal, and its total or partial reprinting must be authorized by SOBRAEP.

Figures, tables and equations should follow the following guidelines.

B. Figures and Tables

Tables and figures (drawings or pictures) should be inserted in the text right after they are mentioned for the first time, as long as they fit the size of the columns; if necessary, use the whole page. Figures resolution should be at least 300 dpi and vector files should be preferably used for better print quality. Table captions should be above the tables and figure captions should be below the figures. The tables should have titles and they are designated by the word Table, being numbered in sequence by Roman numerals. Table captions must be centered and in bold.

Figures also need captions and they are designated by Figure in the text (Fig. in the caption itself), numbered with Arabic numerals in a sequenced manner, left- and right-justified, as shown in the example. The designation of the parts of a figure is done by adding lowercase letters to the numbers of the figures starting with the letter a, e.g. Figure 1(a).

Fig. 4. Magnetization as a function of applied field. (Note that "Fig." is abbreviated and there is a period after the figure number followed by two spaces.)

To better understand graphs, the definition of their axes should be done with words not letters, except when referring to waveforms and phase planes. The units should be between parentheses. For example, use the denomination "Magnetization (A/m)", instead of "M (A/m)".

Figures and tables should be positioned preferably in the beginning or the end of the column, avoiding putting them in the middle. Avoid tables and figures whose sizes exceed the size of the columns. The figures should preferentially be in black, with a white background, since the printed version of the journal is in black and white. Their lines should be thick, so the impression is readable.

C. Abbreviations and Acronyms

Abbreviations and acronyms must be defined the first time they are used in the text, e.g. "... Pulse-Width Modulation (PWM)...".

D. Equations

Number equations consecutively with equation numbers in parentheses flush with the right margin, as in (1). The equations should be written in a compact form, centered in the column. If a nomenclature section is not included in the beginning of the text, the quantities should be defined right after the equation, such as:

$$\Delta I_L = I_o + \frac{\sqrt{3}}{2}\frac{V_i}{Z} \tag{1}$$

where:

ΔI_L - resonant inductor peak current;
I_o - load current;
V_i - source voltage;
Z - characteristic impedance.

VIII. CONCLUSIONS

This paper was fully written in accordance with the guidelines for submissions of papers in English.

ACKNOWLEDGEMENTS

The authors thank John Doe for the collaboration of preparing this paper. This Project was financed by the CNPq (xxyyzz process).

REFERENCES

[1] R. Manhaeve, A. Kimmig, S. Dumančić, T. Demeester, L. De Raedt, "Deepproblog: Neural probabilistic logic programming", *in Advances in Neural Information Processing Systems*, vol. 2018-Decem, pp. 3749–3759, jul 2018, doi:10.48550/arxiv.1907.08194, URL: `https://arxiv.org/abs/1907.08194v2`, `1805.10872`.

[2] J. Johnson, L. Fei-Fei, B. Hariharan, C. L. Zitnick, L. Van Der Maaten, R. Girshick, "CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning", *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 1988–1997, 2017, doi: 10.1109/CVPR.2017.215, `1612.06890`.

[3] K. Heecheol, "kimhc6028/relational-networks: Pytorch implementation of "A simple neural network module for relational reasoning" (Relational Networks)", URL: `https://github.com/kimhc6028/relational-networks`.

[4] R. Manhaeve, "ML-KULeuven/deepproblog: DeepProbLog is an extension of ProbLog that integrates Probabilistic Logic Programming with deep learning by introducing the neural predicate.", URL: `https://github.com/ML-KULeuven/deepproblog`.