Towards Open-Ended VQA Models Using Transformers

Alberto Mario Bellini[1,2] , **Matteo Matteucci**[2] , **Mark James Carman**[1,3] and **Natalie Parde**[2]
[1]Politecnico di Milano
[2]University of Illinois at Chicago
[3]Monash University
albertomario.bellini@mail.polimi.it, matteo.matteucci@polimi.it,
mark.carman@monash.edu, parde@uic.edu

Abstract

In this work, we introduce a new architecture to address the Visual Question Answering problem, an open field of research in the NLP and Vision community. Most of the related work usually share a standard limitation: the number of possible answers is restricted to a limited set of candidates, limiting the power of such models. In this work, we describe a new architecture that employs a new state-of-the-art language model, the Transformer, to generate open-ended answers. In the end, our contribution to the scientific community lies in a new approach that allows VQA systems to generate unconstrained answers. After describing the dataset in use and the architecture itself, we show that our architecture compares well with other VQA models, setting a new baseline for future research.

1 Introduction

Visual Question Answering is an open, multidisciplinary field of research that combines Vision, Natural Language Understanding, and Deep Learning to answer open-domain questions about images. This task is challenging since it involves both the understanding of what is being asked (i.e., the question, asked in natural language) and the reasoning on the associated image to seek relevant information. Current state-of-the-art architectures achieve excellent results and can generalize reasonably well on different question-image pairs. Still, the majority of them share a standard limitation: the generated answers are short and concise, and lie within a fixed set of possible candidates. This is due to how the problem is addressed: a final fully-connected layer is usually employed to distribute probabilities over a restricted vocabulary of answers. [TODO..]

2 Template stuff...

The *IJCAI–20 Proceedings* will be printed from electronic manuscripts submitted by the authors. These must be PDF (*Portable Document Format*) files formatted for 8-1/2$''$ $\times$ 11$''$ paper.

2.1 Length of Papers

All paper *submissions* must have a maximum of six pages, plus at most one for references. The seventh page cannot contain **anything** other than references.

The length rules may change for final camera-ready versions of accepted papers and will differ between tracks. Some tracks may include only references in the last page, whereas others allow for any content in all pages. Similarly, some tracks allow you to buy a few extra pages should you want to, whereas others don't.

If your paper is accepted, please carefully read the notifications you receive, and check the proceedings submission information website[1] to know how many pages you can finally use. That website holds the most up-to-date information regarding paper length limits at all times. Please notice that if your track allows for a special references-only page, the **references-only page(s) cannot contain anything else than references** (i.e.: do not write your acknowledgments on that page or you will be charged for it).

2.2 Word Processing Software

As detailed below, IJCAI has prepared and made available a set of LaTeX macros and a Microsoft Word template for use in formatting your paper. If you are using some other word processing software, please follow the format instructions given below and ensure that your final paper looks as much like this sample as possible.

3 Style and Format

LaTeX and Word style files that implement these instructions can be retrieved electronically. (See Appendix A for instructions on how to obtain these files.)

3.1 Layout

Print manuscripts two columns to a page, in the manner in which these instructions are printed. The exact dimensions for pages are:

- left and right margins: .75$''$
- column width: 3.375$''$
- gap between columns: .25$''$

[1]https://proceedings.ijcai.org/info

- top margin—first page: 1.375″
- top margin—other pages: .75″
- bottom margin: 1.25″
- column height—first page: 6.625″
- column height—other pages: 9″

All measurements assume an 8-1/2″ × 11″ page size. For A4-size paper, use the given top and left margins, column width, height, and gap, and modify the bottom and right margins as necessary.

3.2 Format of Electronic Manuscript

For the production of the electronic manuscript, you must use Adobe's *Portable Document Format* (PDF). A PDF file can be generated, for instance, on Unix systems using `ps2pdf` or on Windows systems using Adobe's Distiller. There is also a website with free software and conversion services: http://www.ps2pdf.com. For reasons of uniformity, use of Adobe's *Times Roman* font is strongly suggested. In LaTeX2e this is accomplished by writing

```
\usepackage{times}
```

in the preamble.[2]

Additionally, it is of utmost importance to specify the **letter** format (corresponding to 8-1/2″ × 11″) when formatting the paper. When working with `dvips`, for instance, one should specify `-t letter`.

3.3 Title and Author Information

Center the title on the entire width of the page in a 14-point bold font. The title must be capitalized using Title Case. Below it, center author name(s) in 12-point bold font. On the following line(s) place the affiliations, each affiliation on its own line using 12-point regular font. Matching between authors and affiliations can be done using numeric superindices. Optionally, a comma-separated list of email addresses follows the affiliation(s) line(s), using 12-point regular font.

Blind Review

In order to make blind reviewing possible, authors must omit their names and affiliations when submitting the paper for review. In place of names and affiliations, provide a list of content areas. When referring to one's own work, use the third person rather than the first person. For example, say, "Previously, Gottlob [1992] has shown that…", rather than, "In our previous work [Gottlob, 1992], we have shown that…" Try to avoid including any information in the body of the paper or references that would identify the authors or their institutions. Such information can be added to the final camera-ready version for publication.

3.4 Abstract

Place the abstract at the beginning of the first column 3″ from the top of the page, unless that does not leave enough room for the title and author information. Use a slightly smaller width than in the body of the paper. Head the abstract with "Abstract" centered above the body of the abstract in a 12-point bold font. The body of the abstract should be in the same font as the body of the paper.

The abstract should be a concise, one-paragraph summary describing the general thesis and conclusion of your paper. A reader should be able to learn the purpose of the paper and the reason for its importance from the abstract. The abstract should be no more than 200 words long.

3.5 Text

The main body of the text immediately follows the abstract. Use 10-point type in a clear, readable font with 1-point leading (10 on 11).

Indent when starting a new paragraph, except after major headings.

3.6 Headings and Sections

When necessary, headings should be used to separate major sections of your paper. (These instructions use many headings to demonstrate their appearance; your paper should have fewer headings.). All headings should be capitalized using Title Case.

Section Headings

Print section headings in 12-point bold type in the style shown in these instructions. Leave a blank space of approximately 10 points above and 4 points below section headings. Number sections with arabic numerals.

Subsection Headings

Print subsection headings in 11-point bold type. Leave a blank space of approximately 8 points above and 3 points below subsection headings. Number subsections with the section number and the subsection number (in arabic numerals) separated by a period.

Subsubsection Headings

Print subsubsection headings in 10-point bold type. Leave a blank space of approximately 6 points above subsubsection headings. Do not number subsubsections.

Titled paragraphs. You should use titled paragraphs if and only if the title covers exactly one paragraph. Such paragraphs should be separated from the preceding content by at least 3pt, and no more than 6pt. The title should be in 10pt bold font and ended with a period. After that, a 1em horizontal space should follow the title before the paragraph's text.

In LaTeX titled paragraphs should be typeset using

```
\paragraph{Title.} text.
```

Acknowledgements

You may include an unnumbered acknowledgments section, including acknowledgments of help from colleagues, financial support, and permission to publish. If present, acknowledgements must be in a dedicated, unnumbered section appearing after all regular sections but before any appendices or references.

Use

```
\section*{Acknowledgements})
```

to typeset the acknowledgements section in LaTeX.

[2] You may want also to use the package `latexsym`, which defines all symbols known from the old LaTeX version.

Appendices

Any appendices directly follow the text and look like sections, except that they are numbered with capital letters instead of arabic numerals. See this document for an example.

References

The references section is headed "References", printed in the same style as a section heading but without a number. A sample list of references is given at the end of these instructions. Use a consistent format for references. The reference list should not include unpublished work.

3.7 Citations

Citations within the text should include the author's last name and the year of publication, for example [Gottlob, 1992]. Append lowercase letters to the year in cases of ambiguity. Treat multiple authors as in the following examples: [Abelson *et al.*, 1985] or [Baumgartner *et al.*, 2001] (for more than two authors) and [Brachman and Schmolze, 1985] (for two authors). If the author portion of a citation is obvious, omit it, e.g., Nebel [2000]. Collapse multiple citations as follows: [Gottlob *et al.*, 2002; Levesque, 1984a].

3.8 Footnotes

Place footnotes at the bottom of the page in a 9-point font. Refer to them with superscript numbers.[3] Separate them from the text by a short line.[4] Avoid footnotes as much as possible; they interrupt the flow of the text.

4 Illustrations

Place all illustrations (figures, drawings, tables, and photographs) throughout the paper at the places where they are first discussed, rather than at the end of the paper.

They should be floated to the top (preferred) or bottom of the page, unless they are an integral part of your narrative flow. When placed at the bottom or top of a page, illustrations may run across both columns, but not when they appear inline.

Illustrations must be rendered electronically or scanned and placed directly in your document. They should be cropped outside latex, otherwise portions of the image could reappear during the post-processing of your paper. All illustrations should be understandable when printed in black and white, albeit you can use colors to enhance them. Line weights should be 1/2-point or thicker. Avoid screens and superimposing type on patterns, as these effects may not reproduce well.

Number illustrations sequentially. Use references of the following form: Figure 1, Table 2, etc. Place illustration numbers and captions under illustrations. Leave a margin of 1/4-inch around the area covered by the illustration and caption. Use 9-point type for captions, labels, and other text in illustrations. Captions should always appear below the illustration.

[3]This is how your footnotes should appear.

[4]Note the line separating these footnotes from the text.

Scenario	δ	Runtime
Paris	0.1s	13.65ms
Paris	0.2s	0.01ms
New York	0.1s	92.50ms
Singapore	0.1s	33.33ms
Singapore	0.2s	23.01ms

Table 1: Latex default table

Scenario	δ (s)	Runtime (ms)
Paris	0.1	13.65
	0.2	0.01
New York	0.1	92.50
Singapore	0.1	33.33
	0.2	23.01

Table 2: Booktabs table

5 Tables

Tables are considered illustrations containing data. Therefore, they should also appear floated to the top (preferably) or bottom of the page, and with the captions below them.

If you are using LaTeX, you should use the `booktabs` package, because it produces better tables than the standard ones. Compare Tables 1 and 2. The latter is clearly more readable for three reasons:

1. The styling is better thanks to using the `booktabs` rulers instead of the default ones.
2. Numeric columns are right-aligned, making it easier to compare the numbers. Make sure to also right-align the corresponding headers, and to use the same precision for all numbers.
3. We avoid unnecessary repetition, both between lines (no need to repeat the scenario name in this case) as well as in the content (units can be shown in the column header).

6 Formulas

IJCAI's two-column format makes it difficult to typeset long formulas. A usual temptation is to reduce the size of the formula by using the `small` or `tiny` sizes. This doesn't work correctly with the current LaTeX versions, breaking the line spacing of the preceding paragraphs and title, as well as the equation number sizes. The following equation demonstrates the effects (notice that this entire paragraph looks badly formatted):

$$x = \prod_{i=1}^{n} \sum_{j=1}^{n} j_i + \prod_{i=1}^{n} \sum_{j=1}^{n} i_j + \prod_{i=1}^{n} \sum_{j=1}^{n} j_i + \prod_{i=1}^{n} \sum_{j=1}^{n} i_j + \prod_{i=1}^{n} \sum_{j=1}^{n} j_i \quad (1)$$

Reducing formula sizes this way is strictly forbidden. We **strongly** recommend authors to split formulas in multiple lines when they don't fit in a single line. This is the easiest approach to typeset those formulas and provides the most

readable output

$$x = \prod_{i=1}^{n}\sum_{j=1}^{n} j_i + \prod_{i=1}^{n}\sum_{j=1}^{n} i_j + \prod_{i=1}^{n}\sum_{j=1}^{n} j_i + \prod_{i=1}^{n}\sum_{j=1}^{n} i_j + + \prod_{i=1}^{n}\sum_{j=1}^{n} j_i \quad (2)$$

If a line is just slightly longer than the column width, you may use the `resizebox` environment on that equation. The result looks better and doesn't interfere with the paragraph's line spacing:

$$x = \prod_{i=1}^{n}\sum_{j=1}^{n} j_i + \prod_{i=1}^{n}\sum_{j=1}^{n} i_j + \prod_{i=1}^{n}\sum_{j=1}^{n} j_i + \prod_{i=1}^{n}\sum_{j=1}^{n} i_j + \prod_{i=1}^{n}\sum_{j=1}^{n} j_i \quad (3)$$

This last solution may have to be adapted if you use different equation environments, but it can generally be made to work. Please notice that in any case:

- Equation numbers must be in the same font and size than the main text (10pt).
- Your formula's main symbols should not be smaller than small text (9pt).

For instance, the formula

$$x = \prod_{i=1}^{n}\sum_{j=1}^{n} j_i + \prod_{i=1}^{n}\sum_{j=1}^{n} i_j + \prod_{i=1}^{n}\sum_{j=1}^{n} j_i + \prod_{i=1}^{n}\sum_{j=1}^{n} i_j + \prod_{i=1}^{n}\sum_{j=1}^{n} j_i + \prod_{i=1}^{n}\sum_{j=1}^{n} i_j \quad (4)$$

would not be acceptable because the text is too small.

7 Examples, Definitions, Theorems and Similar

Examples, definitions, theorems, corollaries and similar must be written in their own paragraph. The paragraph must be separated by at least 2pt and no more than 5pt from the preceding and succeeding paragraphs. They must begin with the kind of item written in 10pt bold font followed by their number (e.g.: Theorem 1), optionally followed by a title/summary between parentheses in non-bold font and ended with a period. After that the main body of the item follows, written in 10 pt italics font (see below for examples).

In LaTeX We strongly recommend you to define environments for your examples, definitions, propositions, lemmas, corollaries and similar. This can be done in your LaTeX preamble using `\newtheorem` – see the source of this document for examples. Numbering for these items must be global, not per-section (e.g.: Theorem 1 instead of Theorem 6.1).

Example 1 (How to write an example)**.** *Examples should be written using the example environment defined in this template.*

Theorem 1. *This is an example of an untitled theorem.*

You may also include a title or description using these environments as shown in the following theorem.

Theorem 2 (A titled theorem)**.** *This is an example of a titled theorem.*

Algorithm 1 Example algorithm

Input: Your algorithm's input
Parameter: Optional list of parameters
Output: Your algorithm's output

```
Let t = 0.
while condition do
    Do some action.
    if conditional then
        Perform task A.
    else
        Perform task B.
    end if
end while
return solution
```

8 Proofs

Proofs must be written in their own paragraph separated by at least 2pt and no more than 5pt from the preceding and succeeding paragraphs. Proof paragraphs should start with the keyword "Proof." in 10pt italics font. After that the proof follows in regular 10pt font. At the end of the proof, an unfilled square symbol (qed) marks the end of the proof.

In LaTeX proofs should be typeset using the `\proof` environment.

Proof. This paragraph is an example of how a proof looks like using the `\proof` environment. □

9 Algorithms and Listings

Algorithms and listings are a special kind of figures. Like all illustrations, they should appear floated to the top (preferably) or bottom of the page. However, their caption should appear in the header, left-justified and enclosed between horizontal lines, as shown in Algorithm 1. The algorithm body should be terminated with another horizontal line. It is up to the authors to decide whether to show line numbers or not, how to format comments, etc.

In LaTeX algorithms may be typeset using the `algorithm` and `algorithmic` packages, but you can also use one of the many other packages for the task.

Acknowledgments

The preparation of these instructions and the LaTeX and BibTeX files that implement them was supported by Schlumberger Palo Alto Research, AT&T Bell Laboratories, and Morgan Kaufmann Publishers. Preparation of the Microsoft Word file was supported by IJCAI. An early version of this document was created by Shirley Jowell and Peter F. Patel-Schneider. It was subsequently modified by Jennifer Ballentine and Thomas Dean, Bernhard Nebel, Daniel Pagenstecher, Kurt Steinkraus, Toby Walsh and Carles Sierra. The current version has been prepared by Marc Pujol-Gonzalez and Francisco Cruz-Mencia.

A LaTeX and Word Style Files

The LaTeX and Word style files are available on the IJCAI–PRICAI–20 website, https://www.ijcai20.org/. These style files implement the formatting instructions in this document.

The LaTeX files are `ijcai20.sty` and `ijcai20.tex`, and the BibTeX files are `named.bst` and `ijcai20.bib`. The LaTeX style file is for version 2e of LaTeX, and the BibTeX style file is for version 0.99c of BibTeX (*not* version 0.98i). The `ijcai20.sty` style differs from the `ijcai19.sty` file used for IJCAI–19.

The Microsoft Word style file consists of a single file, `ijcai20.doc`. This template differs from the one used for IJCAI–19.

These Microsoft Word and LaTeX files contain the source of the present document and may serve as a formatting sample.

Further information on using these styles for the preparation of papers for IJCAI–PRICAI–20 can be obtained by contacting `pcchair@ijcai20.org`.

References

[Abelson *et al.*, 1985] Harold Abelson, Gerald Jay Sussman, and Julie Sussman. *Structure and Interpretation of Computer Programs*. MIT Press, Cambridge, Massachusetts, 1985.

[Baumgartner *et al.*, 2001] Robert Baumgartner, Georg Gottlob, and Sergio Flesca. Visual information extraction with Lixto. In *Proceedings of the 27th International Conference on Very Large Databases*, pages 119–128, Rome, Italy, September 2001. Morgan Kaufmann.

[Brachman and Schmolze, 1985] Ronald J. Brachman and James G. Schmolze. An overview of the KL-ONE knowledge representation system. *Cognitive Science*, 9(2):171–216, April–June 1985.

[Gottlob *et al.*, 2002] Georg Gottlob, Nicola Leone, and Francesco Scarcello. Hypertree decompositions and tractable queries. *Journal of Computer and System Sciences*, 64(3):579–627, May 2002.

[Gottlob, 1992] Georg Gottlob. Complexity results for nonmonotonic logics. *Journal of Logic and Computation*, 2(3):397–425, June 1992.

[Levesque, 1984a] Hector J. Levesque. Foundations of a functional approach to knowledge representation. *Artificial Intelligence*, 23(2):155–212, July 1984.

[Levesque, 1984b] Hector J. Levesque. A logic of implicit and explicit belief. In *Proceedings of the Fourth National Conference on Artificial Intelligence*, pages 198–202, Austin, Texas, August 1984. American Association for Artificial Intelligence.

[Nebel, 2000] Bernhard Nebel. On the compilability and expressive power of propositional planning formalisms. *Journal of Artificial Intelligence Research*, 12:271–315, 2000.