An extended evaluation of the impact of different modules in ST-VQA systems

Viviana Beltrán, Mickaël Coustaty, Nicholas Journet, Juan Caicedo, Antoine Doucet

An extended evaluation of the impact of different modules in ST-VQA systems[*]

Viviana Beltrán[1][0000−0003−1227−719X], Mickaël Coustaty[1][0000−0002−0123−439X], Nicholas Journet[2][0000−0002−6773−4071], Juan C. Caicedo[3,4][0000−0002−1277−4631], and Antoine Doucet[1][0000−0001−6160−3356]

[1] University of La Rochelle, 17000, France
{vbeltran,mickael.coustaty,antoine.doucet}@univ-lr.fr
[2] University of Bordeaux, 33000, France
journet@labri.fr
[3] Fundación Universitaria Konrad Lorenz, Bogotá, Colombia
[4] Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
jcaicedo@broadinstitute.org

Abstract. Scene Text VQA has been recently proposed as a new challenging task in the context of multimodal content description. The aim is to teach traditional VQA models to read text contained in natural images by performing a semantic analysis between the visual content and the textual information contained in associated questions to give the correct answer. In this work, we present results obtained after evaluating the relevance of different modules in the proposed frameworks using several experimental setups and baselines, as well as to expose some of the main drawbacks and difficulties when facing this problem. We makes use of a strong VQA architecture and explore key model components such as suitable embeddings for each modality, relevance of the dimension of the answer space, calculation of scores and appropriate selection of the number of spaces in the copy module, and the gain in improvement when additional data is sent to the system. We make emphasis and present alternative solutions to the out-of-vocabulary (OOV) problem which is one of the critical points when solving this task. For the experimental phase, we make use of the TextVQA database, which is one of the main databases targeting this problem.

Keywords: Visual Question Answering · Scene Text Recognition · Deep Learning · Copy Module.

1 Introduction

Scene Text Visual Question Answering (ST-VQA) targets the specific task where understanding the textual information in a scene (text contained in signs, posters or ads in the image) is required in order to give the correct answer. Although

[*] This work has been supported by the French region of Nouvelle Aquitaine under the Animons project and by the Mires research federation.

deep learning has been used with acceptable accuracy results of ≈ *70%* for traditional VQA tasks, when solving the ST-VQA problem the accuracy drops to ≈ *27%* demonstrating the challenge ahead. As a semantic description task, many related issues need to be addressed, we mention the most relevant ones. First, understand the type of question: although the task is well defined, current databases fail to define clean samples fitting the specifications of the task. Fig. 1 presents samples of triplets *(Image, Question and Answers)* from TextVQA database [13]. Each sample contains an image, a question associated and the ground truth list of answers given for 10 human annotators (for cases where the answers given by annotators are all the same, we put the unique answer, if the answers are different, we put the entire set of answers given by annotators). We present examples for wrong annotation cases *(A-E)* where there is no need to read the text in the image to answer the question or the answer given by the annotator is not correct (*case C*), the answer fall into the category 'Yes'/'No' *(case D)* or the case where the sample doesn't fit the requirements but it is possible to filter those samples by the answer given by the annotator *(case E)*, and correct cases when the answer is as expected, some text present in the image required in the question *(case F)*.

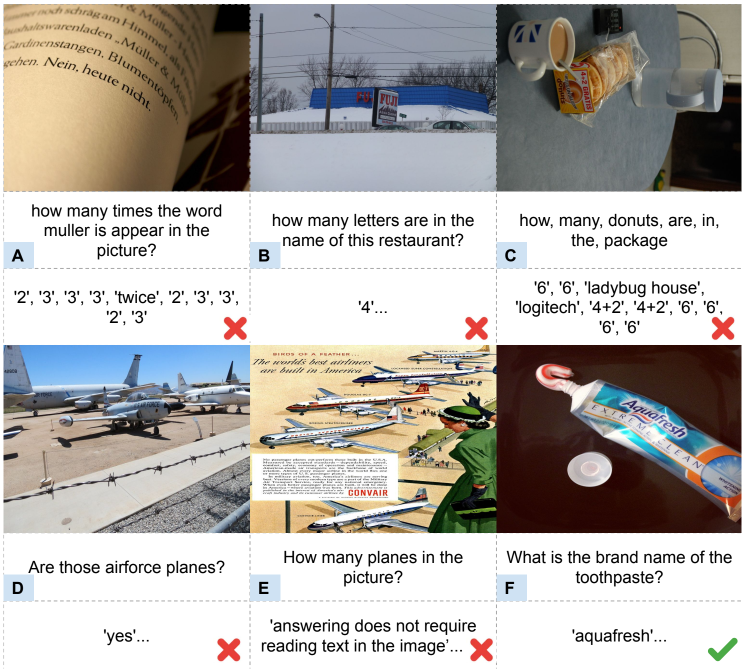
A
how many times the word muller is appear in the picture?
'2', '3', '3', '3', 'twice', '2', '3', '3', '2', '3'
✗
B
how many letters are in the name of this restaurant?
'4'...
✗
C
how, many, donuts, are, in, the, package
'6', '6', 'ladybug house', 'logitech', '4+2', '4+2', '6', '6', '6', '6'
✗
D
Are those airforce planes?
'yes'...
✗
E
How many planes in the picture?
'answering does not require reading text in the image'...
✗
F
What is the brand name of the toothpaste?
'aquafresh'...
✓

Fig. 1. Samples of triplets (Image, Question, List of answers given by 10 annotators) from TextVQA database for wrong (A-E) and correct (F) annotation cases for the problem of ST-VQA.

Other problem associated, is the detection and recognition of the text present in the visual data in the wild or from natural scenes, which remains a challenge for current models facing this problem because of all variations contained in them [9]. One of the most difficult tasks, even in traditional VQA systems [6] is the reasoning required to resolve spatial and visual references that involves understanding about the question and the visual information at the same time. Another related problem is the representation of the answer space, as this can contain unlimited words in any possibly language, which makes infeasible the establishment of a fixed pool of answers, and yields the problem of "out-of-vocabulary" answers (words not contained in the pool of answers are often called OOV words).

In this document, we present an incremental ablation study and an analysis for the modules comprised when solving the ST-VQA task. Our contributions are the following: we evaluate a strong architecture widely used in the context of media description for VQA systems applied to the problem of ST-VQA. We tested representative feature extractor models for the modalities involved in the VQA system. We also evaluate the relevance of the dimension of the answer space for the case of fixed set of words and for the case when the copy module is used as the main strategy for the OOV problem. We expose drawbacks related with the copy module which is the state-of-the-art solution, as well as the proposal of using a second metric to compute the scores for the dynamic spaces so that the copy module can take advantage of texts not 100% recognized for the OCR system. We evaluate the performance of including additional data to train the system in the form of a complementary network representing embeddings from textual and visual data. Finally, we present the results of several ablative studies to validate the relevance of the proposed analysis, making use of TextVQA database with baseline results in the validation set to facilitate performance evaluation.

2 Related work

The very first work introducing the task was proposed by Singh et al. [13], they created a new database called TextVQA and presented a strategy (LoRRA) based on deep learning to solve the task. Their strategy contains the following components, a VQA system to process inputs obtained by using an object detection model for the visual features and GloVe vectors [11] to encode the question, a reading component to include OCRs extracted by using a text recognition model as weighted Fasttext features [5] and an answering module composed of a fixed + a dynamic answer space included by using a copy module (see Section 3.4) to handle the OOV problem which is one of the biggest problems to address in this task.

The second most relevant work at the time of writing this article, was presented by Biten et al. [1]. Similar to [13], they also introduced a new database (we refer to this database as ICDAR db) created for the ICDAR 2019 Robust Reading Challenge on Scene Text Visual Question Answering (see [4] for details on the competition). In this work, they presented final results for the proposed competitions from different participants addressing the task of ST-VQA under three tasks of increasing difficulty. The winning strategy, VTA, makes use of a strong architecture based on two types of attention, bottom-up (since they used an object detection model as a visual feature extractor) and top-down attention (by including OCR information extracted with an OCR recognition system). For the text, they use a pre-trained Bert model [14] to turn all the text into sentence embeddings, including object names, OCR recognition results, questions and answers (from training set). Having these embeddings for the text and for the images, they use a similar architecture as the one presented by Anderson et al. [2] to get the answer.

By analysing models and results presented for both main strategies, LoRRA [13] and VTA [4], their architectures are very similar, as well as obtained results in the target database, TextVQA and ICDAR db respectively. They both use an object detection model to extract visual features, an embedding method for the text (question, answers, OCRs), a VQA system, and an answer module. For VTA, the accuracy in task 1 of the challenge is *43.52%* which is significant better than the accuracy for tasks 2 *(17.77%)* and 3 *(18.13%)*, (for the complete definition of the tasks, see "Table II: Main Results Table" at [1], where tasks are 1: Strongly Contextualized, 2: Weakly Contextualise, and 3: Open Dictionary). In task 1, ground truth text is provided in the database as a set of possible words related to the scene, while for tasks 2 and 3, these texts are obtained by using an OCR recognition system and therefore relying on its performance to obtain ground truth texts from the images. By making a fair comparison, the results obtained in tasks 2 and 3 can be compared to LoRRA, with an accuracy of *27.63%* (see "Table 2: Evaluation on TextVQA" at [13]), as in both cases, the strategy relies on using OCRs obtained for a recognition system. The strategy in [13] performs better because of the inclusion of the copy module (see Section 3.4) that allow to handle the OOV problem, however, as we will discuss in Section 5, this solution is far from being optimal as it presents many limitations and a dubious performance.

Other works such as [3] describe the strategy used for one of the competitors, VQA-DML in [4], in which the main difference is the use of a n-gram representation for the answer space that allow to handle the OOV problem as well as giving the system the possibility to extend the answer space but not the dimension of it (i.e., the number of possible words formed from a n-gram combination increases, while the dimension of the target vector keeps a reasonable size). However, it also poses another challenge as it is required to add an additional stage for retrieving the correct answer from the n-gram predicted representation. While the low accuracy reported of VQA-DML for ICDAR db database (approx. 11%) [4] can be attributed to the straightforward architecture used for their authors, more analysis are required to determine the convenience of using this n-gram representation for the answer space. As this task is attracting attention, recent works present the task by introducing new databases, [10] introduces a new database, OCR-VQA–200K comprising images of bookcovers, [12] introduces a database containing images of business brands, movie posters and book covers.

3 Architecture description

In order to perform comparisons with the state-of-the-art, we make use of similar frameworks for our experimental setups. Taking into account Fig. 2, we can divide the framework into modules. The following is the description of each one.

3.1 The embeddings module

The embeddings module represents the process of computing input features for the modalities involved, visual and textual (and other possible data such as OCRs, and localized features). For this reason, different specialized models for each modality can be studied (see Section 4.3 for information of the models tested during experimentation phase).

3.2 VQA model

The VQA module represents the component in which the data is combined. We make use of a similar architecture as the one presented in [2]. It inputs the features extracted by using module A and used them to train the network and give the correct answer. We make use of attention mechanisms directed from the question network to the visual network (and the complementary network, see Table 3) .

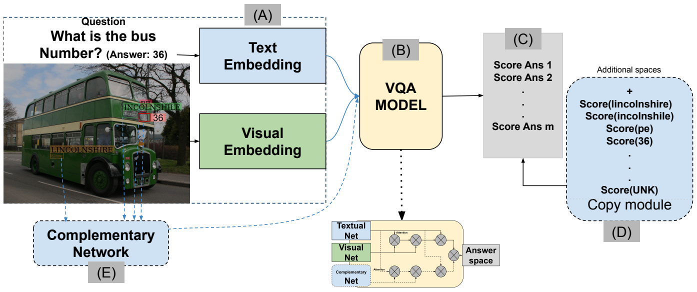

Fig. 2. Modules comprised in ST-VQA frameworks. Modules A, B and C represent the basic modules comprised in an STVQA framework. Modules D and E are added as strategies for improve the performance. A) Embedding module for input data of different modalities (Visual/Textual/OCRs), B) VQA system, C) Answer space, D) Copy Module, and E) Complementary network.

3.3 Answer space module

The answer module is in charge of the representation of the target vector relying on an answer space. We evaluated the usage of a fixed answer space commonly know as a bag of words (BoW), in which the score of each space in the final vector will indicate the presence or absence of the word.

3.4 Copy Module

The copy module works as a mechanism to handle the OOV problem. For this task, it is specially required, because the dimension of the answer space can growth unlimited. It works by adding a set of additional spaces to the fixed answer space (module C), filled with scores computed by using the OCRs recognized in the image. Thus, the final dimension of the answer space will be the one fixed by the set of selected answers from the training data + the set of dynamic words with a fixed number of spaces representing the OCRs.

We propose to compute the scores using two different metrics. First, the *Human Score* metric used in [13] computed as follows:

$$HS(\mathbf{ans}) = min(\frac{\#\, humans\, that\, said\, \mathbf{ans}}{3}, 1) \tag{1}$$

Each OCR will be taken as the *'ans'* to compute the score, this means that for *'ans'* to get a $HS = 1$, *'ans'* should be present in the set of answers given by the annotators at least three times. Fig. 3 shows an example of calculation of scores using eq. 1 for an image with two different questions associated. For the first question Qi, the answer is composed of two words, (*eddie, izzard*), which are outside the fixed answer space. The copy module could help to use the OCRs extracted from the image as an advantage, however, in this case, the *Human Score, eq. 1*, will be zero for all the OCRs, because it seeks a perfect match between the ground truth answer *'eddie izzard'* and each one of the OCRs in a separate way *['eddie', 'izzard']* , leading to a zero vector as the target representation for this sample. For the second question associated Qj, it works as expected, as there exists an exact match between the ground truth answer and the OCRs.

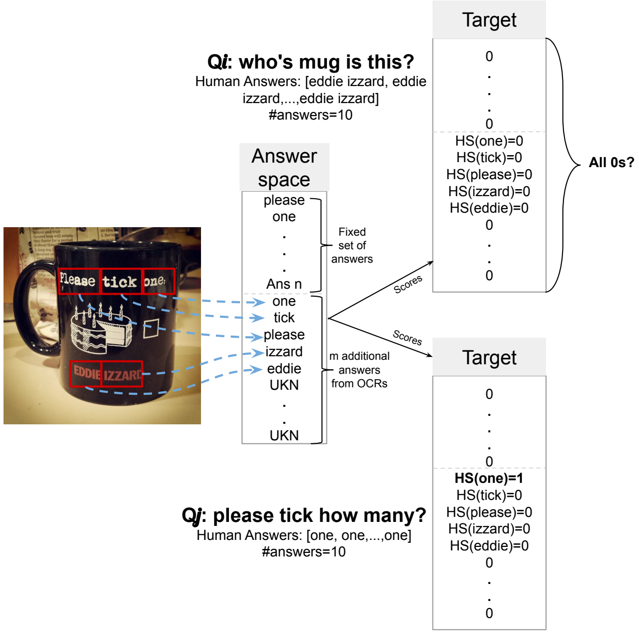

Fig. 3. Sample of assignation of scores using Human Score in the positive and negative cases of match between ground truth answers and the set of OCRs of the image.

Another example of the use of the copy module is when the set of human answers contains more than one answer for the same question, as it is expected to be the same for all the 10 annotators, but in some cases, they can differ and give a different answer. As an example, if the set of human answers is *[stop, emergency stop, emergency stop, stop, stop, emergency stop, emergency stop, it is an emergency stop, emergency stop, unanswerable]*, and the set of OCRs recognized is *[stop, emergency]*, although the majority of answers in the ground truth is *’emergency stop’*, it also contains another ground truth answers such as *’stop’*, which results convenient at setting a score for the OCR *’stop’*. As these two previous cases, there are others in which the copy module may or may not work because it relies in having texts 100% well recognized in the images (*’one’ is completely different from ’one:’ when computing the Human Score*), or when answers contain more than one token.

In order to improve the calculation of scores when partial matches are found in the OCRs of the image, we propose to use a second metric based on the Average Normalized Levenshtein Similarity (ANLS), computed as follows:

$$ANLS\ Score(ocr) = \frac{1}{M}\sum_{i=0}^{M-1} 1 - NL(ans_i, ocr)) \tag{2}$$

where M = 10 is the set of answers given by 10 human annotators. Thus, for the previous example in Fig. 3, for *Qi*, the new scores for the target OCRs will be *Score("izzard") = 0.4166* and *Score("eddie") = 0.5*

There are also problems related with the OCR system used, such as recognizing a single word separated in each one of their characters, or non recognizing the target text (i.e., the ground truth word from the human answer set that truly appears in the image). In Section 5, we discussed about the advantages and disadvantages for training models that uses the copy module as the main strategy to solve the OOV problem, as well as the dubious results when evaluating the performance in validation.

3.5 Complementary Network

The module E represents the inclusion of networks that input additional data into the VQA system. We tested three different setups: First, Fasttext embeddings [5] from OCRs recognized in the images (we use the OCRs available in the database) as in [13], however, we do not concatenate the order of the OCRs, and we input the average of the embeddings from all the OCRs available without weighting them. Second, we use global features extracted from the visual boxes containing the text recognized in the image, for training, we use the ground truth answers to filter the boxes with text matching in at least 30% one of the answers, for validation, we use the entire image. We also test the scenario when both, Fasttext and global features are sent into the VQA system together.

4 Experiments

4.1 Databases

As we mention in section 2, in order to be able to explore different evaluation scenarios, we are using the TextVQA database [13], as it provides baselines results for the validation set for comparison purposes. This database contains 34,602 training samples and 5,000 validation samples, with almost 50% of the answers being unique. This shows the difficulty of using a fixed set of words in the answer space.

4.2 Evaluation metrics

Two main metrics are used to evaluate the performance in this task: First, the accuracy computed as follows:

$$accuracy(y, \hat{y}) = \frac{1}{N_{samples}} \sum_{k=0}^{N_{samples}-1} 1(\hat{y}_i = y_i) \tag{3}$$

where $1(k)$ is the indicator function.

In the cases where the copy module is used, the calculation of the accuracy changes by using the Human Score accuracy (Eq. 1), the predicted answer is obtained by getting the index of the max value in the output vector, if the index is in the first part of the prediction (fixed space), the answer will be one of the fixed space shared among all the samples, if the index is in the additional / dynamic part of the output vector, the answer will be one of the set of OCRs recognized in the image at the index position. The second performance metric is the Average Normalized Levenshtein Similarity (ANLS) computed as follows:

$$ANLS = \frac{1}{N} \sum_{i=0}^{N} \left(max_j \, s(a_{ij}, o_{qi}) \right) \tag{4}$$

$s(a_{ij}, o_{qi}) = \begin{Bmatrix} (1 - NL(a_{ij}, o_{qi})) \; if \, NL(a_{ij}, o_{qi}) < \tau \\ 0 \qquad\qquad if \, NL(a_{ij}, o_{qi}) \geq \tau \end{Bmatrix}$ where N is the total number of questions, M is the total number of ground truth answers per question, a_{ij} the ground truth answers where i = 0, ..., N, and j = 0, ..., M, o_{qi} be the network's answer for the i-th question q_i, and τ is the threshold that determines if the answer has been correctly selected but not properly recognized, or on the contrary, the output is a wrong text selected from the options and given as an answer [4].

For this task, the second metric, ANLS (eq. 4), could be more convenient, as the system can find partial matches among the set of words in the answer space. On the contrary, evaluating the accuracy (eq. 3), imposes a huge penalty if the model does not find perfect matches for the answers. This is directly related with the answer module and the OOV strategy used. However, taking into account the value of τ for the calculation of ANLS score that penalizes predictions matching in less than 50% of the characters, the performance for both metrics is expected to be similar.

4.3 Baselines and Ablations

We perform several ablation studies for the modules described in Section 3, where we aim to analyse the performance, drawbacks and future improvements when targeting this task. We describe the ablations performed, to see if adding/ changing/ replacing key modules of the system would lead us to obtain better results. For the first 5 models, we wanted to analyse the embedding module (see module A of Fig. 2), for the image and question data, and select the best one to test the rest of evaluation scenarios. The components included in this set of studies from Fig. 2 are modules A, B and C. For the answer space, we use the set of 3997 most frequent answers (Small Set SS, where the answers selected are those with frequencies $\geqslant 2$) in the training database. As representative embeddings, we compare two models for the images, ResNet101 [8] and Faster R-CNN (bottom-up (BU) attention) [2] with final representations of *2048-dim* and *36 (features per image with a 2048-dim each one)* respectively. And two embedding models for the text, GloVe [11] and BERT [7], with final representations of *300-dim* and *768-dim* respectively, for both models, we use a set of 15 tokens as the maximum length. The vocabulary size extracted of the questions is 9312 unique words. Therefore, the scenarios evaluated are: GloVe + ResNet, GloVe + BU, BERT + ResNet, BERT + BU.

After selecting the best set of embeddings based on the previous results, i.e., GloVe embeddings for the question, bottom-up features (BU) for the images and having fixed a small answer space (SS), we wanted to evaluate if the performance improve by increasing the size of the answer space to a large one (LS). The last row in Table 1 presents the result by using a larger set for the answer space of 7999 most frequent answers in the training database. Table 1 presents the results obtained for these first 5 models for validation samples in which the answer is contained in the selected fixed set of answers, i.e., for the answer space SS, the # of samples get reduced to 18,516 for training and 2,214 samples in validation. For the answer space LS, the # of samples get reduced to 21,183 for training and 2,290 samples in validation. Thus, to make a fair comparison, results are reported over these validation subsets.

Table 1. Performance for representative embedding models for visual and textual data in ST-VQA systems, with a fixed set of words in the answer space. Validation results are reported over the set of samples which answers are contained in a small fixed set SS or a larger set LS.

Model	Acc	ANLS	AVG
GloVe + ResNet101 + SS	0.1853	0.2274	0.2065
GloVe + BU + SS	**0.2005**	**0.2474**	**0.2240**
BERT + ResNet101 + SS	0.1910	0.2319	0.2115
BERT + BU + SS	0.1978	0.2366	0.2172
GloVe + BU + LS	0.1860	0.2279	0.2069

The second set of evaluation scenarios aim to analyse the inclusion of the copy module, based on results from Table 1. The components included from Fig. 2 are modules A, B, C and D. We wanted to evaluate the appropriate number of additional spaces, for this, we test three different numbers, first, 50 spaces following the work [13], second, by taking the average of OCRs of all training samples (≈ 9.8) * 2, i.e., 20 spaces, and finally, by taking the average of OCRs from all training samples, i.e., 10 spaces. In this case, the data sets contain 100% of samples (34602 for training and 5000 for validation).

As we discussed in Section 3.4, the assignation of scores using equation 1, does not take advantage of text not 100% recognized, leaving many samples with zero score vectors. In this case, we wanted to change the assignation of scores by using the average ANLS score (see equation 2) over the set of human answers. The last row in Table 2 changes the assignation of scores using ANLS score metric. Table 2 presents the results for this set of evaluation scenarios.

Table 2. ST-VQA performance with the inclusion of the copy module with the assignation of scores using Human Score metric and by exploring the number of additional spaces for the OCRs to 50, 20 and 10. The last result changes the calculation of scores using the ANLS score metric.

Model	Acc	ANLS	AVG
50 spaces + Human Score	**0.1854**	**0.1835**	**0.1844**
20 spaces + Human Score	0.1778	0.1799	0.1788
10 spaces + Human Score	0.1792	0.1817	0.1804
50 spaces + ANLS Score	0.1705	0.1816	0.1761

To evaluate if the inclusion of more information into the VQA system could help the performance, we test the inclusion of three complementary data: the average of fasttext embeddings [5] from OCRs recognized in the images, similar as in [13], but without the addition of order and weighted information, a concatenation of global descriptors extracted from boxes containing target text, and finally, by sending into the VQA module both of them. For this evaluation scenario, the components included from Fig. 2 are modules A, B, C, D and E. We use the best model from Table 2, adding a top-down attention in the VQA system from the question towards the complementary network data. Table 3 presents the results obtained for this set of experiments.

Table 3. ST-VQA performance when complementary data is sent into the VQA module. Three types of complementary data were evaluated, Fasttext embeddings from OCRs recognized in the image, Global features extracted from the box containing the target text data and, finally, the combination of them.

Model	Acc	ANLS	AVG
OCR Fasttext	**0.1848**	**0.1942**	**0.1895**
Global Visual features (GVF)	0.1756	0.1797	0.1776
OCR Fasttext + GVF	0.1843	0.1932	0.1887

5 Discussion

The best results from Table 1 are obtained by using GloVe vectors + Fast R-CNN (or bottom-up BU) features. The slightly better performance of GloVe over BERT can be attributed to the fact that the structure and meaning of the words in the questions for this database is shared, and therefore the context does not play an important role in the discrimination of different samples. Also, as the last result in the Table showed, increasing the set of possible answers not necessarily implies an improvement of the performance (see also results of small set SA vs large set LA at "Table 2: Evaluation on TextVQA" [13] that confirm our result). This is because the set of possible answers can contain any combination of characters in different languages that are found in natural images, in the case of TextVQA database, there are more than 19,000 different answers among 34,000 samples. This makes unfeasible the establishment of a manageable fixed set of words as the answer space, and raises the question in how to handle the OOV problem?

For Tables 2 and 3, the copy module was included as a strategy to handle the OOV problem. The best number of additional spaces to include in the answer space for this database was 50, this means that the performance improves as more text data recognized in the image is sent into the system. On the contrary, the last result in the Table that tested the new metric to compute the scores, ANLS Score, did not show an improvement in the performance, which is related to the fact that for the majority of samples in the database with at least one OCR recognized, the answers are composed of only one token, and therefore the scores will be similar (for only 8.9% of the samples in TextVQA, answers contain more than one token).

Finally, regarding the evaluation of the inclusion of additional data, fasttext embeddings showed a small improvement for the performance. On the contrary, the inclusion of the global visual descriptors with the target textual data did not show any relevance, this could be the attention mechanism used, as it is the same used for both embeddings. However, a deeper analysis regarding the optimal attention mechanism is required to determine if the extra data is helping the system to learn, we leave it as a future work.

Is the copy mechanism solving the OOV problem in a suitable way? We wanted to give final comments regarding the convenience of using the copy module as a strategy for the OOV problem. Although, the copy module partially solves the OOV problem, each item in the dynamic space could represent as many different words exist in the OCR space of all samples, and at the end, the prediction of the correct answer over these values becomes almost a randomly choice that depends on the position of the OCR. Better solutions to handle the OOV are required as many tasks in the state-of-the-art are facing the same problem. The n-gram representation for the answer space could be a solution as with this, a larger set of answers can be represented by a fixed and manageable set of n-grams. However, it is required to perform deeper analysis of the implications of its usage.

6 Conclusions

We presented an incremental and extended study for the task of ST-VQA by performing an analysis of the modules required in any framework addressing this task. As one of the main analysed aspects, we evaluated the relevance of the dimension when a fixed set of words (BoW solution) is used as the answer space that for this problem turned out to be of little importance. We also evaluated the performance of the model when using the copy module under two different metrics for the calculation of the scores, both of them ended up with similar performance as the majority of data contains answers with only one token. Our final evaluation was the performance when including complementary data to train the system in the form of an additional network resulting in a slightly improvement of the performance. Finally, we expose some of the main drawbacks of current solutions, specially when handling the OOV problem showing us the need for better and more robust strategies. As a future work, we want to explore the performance when more data is used in the training phase, as we have noticed in the state-of-the-art, data augmentation has not been used when addressing this task. We also want to explore robust OOV strategies that do not rely on the copy mechanism, and finally, to study the mechanisms of inclusion of complementary data into the system that can help in the improvement of the performance.

References

1. Scene text visual question answering. arXiv preprint arXiv:1905.13648 (2019)
2. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6077–6086 (2018)
3. Beltr, V., Journet, N., Coustaty, M., Doucet, A., et al.: Semantic text recognition via visual question answering. In: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). vol. 5, pp. 97–102. IEEE (2019)
4. Biten, A.F., Tito, R., Mafla, A., Gomez, L., Rusiñol, M., Mathew, M., Jawahar, C., Valveny, E., Karatzas, D.: Icdar 2019 competition on scene text visual question answering. arXiv preprint arXiv:1907.00490 (2019)
5. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)
6. Cadene, R., Ben-Younes, H., Cord, M., Thome, N.: Murel: Multimodal relational reasoning for visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1989–1998 (2019)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. Liu, X., Meng, G., Pan, C.: Scene text detection and recognition with advances in deep learning: a survey. International Journal on Document Analysis and Recognition (IJDAR) **22**(2), 143–162 (2019)
10. Mishra, A., Shekhar, S., Singh, A.K., Chakraborty, A.: Ocr-vqa: Visual question answering by reading text in images. ICDAR (2019)
11. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
12. Singh, A.K., Mishra, A., Shekhar, S., Chakraborty, A.: From strings to things: Knowledge-enabled vqa model that can read and reason. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4602–4612 (2019)

13. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. arXiv preprint arXiv:1904.08920 (2019)
14. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Brew, J.: Huggingface's transformers: State-of-the-art natural language processing. ArXiv **abs/1910.03771** (2019)