# QAlayout: Question Answering Layout based on multimodal Attention for visual question answering on corporate Document

Ibrahim Souleiman Mahamoud, Mickaël Coustaty, Aurélie Joseph, Vincent Poulain d'Andecy, Jean-Marc Ogier

# QAlayout: Question Answering Layout based on multimodal Attention for visual question answering on corporate Document

Ibrahim Souleiman Mahamoud (✉)*+,Mickaël Coustaty+
Aurélie Joseph*, Vincent Poulain d'Andecy*, Jean-Marc Ogier+

+ La Rochelle Université, L3i Avenue Michel Crépeau, 17042 La Rochelle, France
Email: {ibrahim.souleiman_mahamoud,mickael.coustaty,jean-marc.ogier}@univ-lr.fr
* Yooz 1 Rue Fleming, 17000 La Rochelle, France
Email: {aurelie.joseph,vincent.poulaindandecy}@getyooz.com

**Abstract.** The extraction of information from corporate documents is increasing in the research field both for its economic aspect and a scientific challenge. To extract this information the use of textual and visual content becomes unavoidable to understand the inherent information of the image. The information to be extracted is most often fixed beforehand (i.e. classification of words by date, total amount, etc). The information to be extracted is evolving, so we would not like to be restricted to predefine word classes . We would like to question a document such as "which is the address of invoicing?" as we can have several addresses in an invoice. We formulate our request as a question and our model will try to answer. Our model got the result 77.65% on the Docvqa dataset while drastically reducing the number of model parameters to allow us to use it in an industrial context and we use an attention model using several modalities that help us in the interpertation of the results obtained. Our other contribution in this paper is a new dataset for Visual Question answering on corporate document of invoices from RVL-CDIP[8]. The public data on corporate documents are less present in the state-of-the-art, this contribution allow us to test our models to the invoice data with the VQA methods.

**Keywords:** Visual question answering · Multimodality · Attention mechanism.

## 1 Introduction

Imagine a near future where you will not process any document but your digital clone will do it for you. It will summarize and/or extract the relevant and useful information for you. For some, this future seems close while for others it is a simple illusion. Some companies have already started this transition such as the New Zealand company UneeQ[16], which created an avatar of Albert Einstein that you can interact with on various topics.

If we go back to the present, companies try to provide easy-to-use solutions to extract information from the corporate documents. These corporate documents

are varied in both content and form (i.e. invoices, order form, resume, pay slip etc). Thus the customer processing are evolving everyday, and if they today only focus on the extraction of information from invoice, tomorrow they will ask for more and more intelligent process of documents (i.e for instance the automatic processing of collaborators resume, or the automatic linking of file folders). We would therefore like to have a method to extract information with few data and able to adapt to another type of document while taking into account the changes in layout and content to be extracted.

To answer these problems, various state-of-the-art papers have tried to provide a solution. Some few-shot learning [19] methods were trained only a corpus composed of a hundred of documents. These methods have all shown their weakness when used with a large dataset. Some other methods, based on deep-learning techniques such as Lambert[6] or LayoutLM[22],have appeared recently. Their performances were impressive but with the drawback that these methods require a lot of data to converge. A very large dataset of annotated documents is not always available and in order to deal with this limitation, some recent works proposed to use incremental methods [18] which can evolve over time.

We propose in this paper a method to process different types of documents and can also extract the different information by the customer for a question answering model. The usefulness of visual question answering on corporate documents unlike predefined extraction is to allow to extract more general information that can be adapted to a new corpus of data.

In general when we read a paper whether it is scientific or other, some questions come to mind. For example for our paper, if you try to understand our contribution you quickly read the abstract or propose method to have an answer. This natural ability to focus on a part of the information from questions is innate to us we look at what is essential to solve a problem. It is by being inspired by this that the mechanism of attention has seen the day several papers such as [17] are sold the merit of mechanism of attention. The attention mechanism is used in several situations, some to optimize the performance of their model, others to be able to interpret the model and thus brings some element of answer to the behavior of this black box. The visual question answering is known to be very popular in the community because many problems can be solved with it, from simple questions on images of natural scenes [7] to medical assistance [11] to help specialties to better understand some images thanks to the strength of accumulating a large amount of information to synthesize and then to keep it in its memory of artificial neural networks.

The visual questions answering approaches on corporate documents are minimal and the data also concerns this subject. In this article we will describe our method based on the Qanet architecture [23]. The key motivation behind the design of our model is as follows: The convolution layer helps to capture the local structure of the image and text. The co-attention layer allows to have a global relationship between the inputs in order to define the positive or negative impact of one in relation to the other. We have evaluated our model on the Squad [14] and Docvqa [13] datasets. The choice of these two datasets is that one is only

textual, it will allow us to test and optimize the textual part of our model while the Docvqa is a corpus of documents of different type, it will allow us to test if our model will manage to use well this multimodal corpus.

The contributions of this paper are summarized as follows:

- We have made available a corpus of Visual question answering data VQA-CD for corporate documents, this corpus is to our knowledge the first in the state-of-the-art. We have annotated 3 thousand questions from $\sim$ 693 documents.
- We propose a multi-modal co-attention model for visual question answering. This template will use the visual and textual content features of this document and the layout features for each word. This model learns the best way to use cross-modality to predict the answer of a question.We use this self-attention to focus our network on common features from the input. This will allow us to exploit the context and query correlation at the initial stage.

## 2 Related work

The approaches proposed for VQA are generally distinguished in three categories. Some use a single modality like Qanet [23]. Although these methods show good performances by using some transformer-based model [17], they remain less effective when the multimodal understanding is necessary.

The second category of methods proposes to rely on multimodal architectures in order to be able to deal with the visual and the textual content at the same time [20]. These methods have then be designed to include a second modality in the proposed architecture. Even if only these models require better performances, this can only be done by using large dataset and they require to re-train the model each time new kind of input is used. This relies on the fact that the document layout may vary in a significant way and this is not taken into account. For example, if you train a model on a dataset where the addresses are generally located at the top of documents, this one will have difficulties to find them elsewhere (addresses could be located at the bottom, or in the middle of new documents and dataset). Such kind of models then need to be re-trained on new dataset which should come with its annotations.

To overcome this problem, some recent models like LayoutLM [22] have been introduced to be able to add a third modality representing the layout information of documents. The most recent and popular model from this third category includes the famous LayoutLM [22], LamBERT [6] and ViBERTgrid [10]. LAMBERT [6] proposed a model based on the Transformer encoder architecture RoBERTa [12]. The main contribution of this paper was to propose a general-purpose language model that views text not simply as a sequence of words, but as a collection of tokens on a two-dimensional page by applying relative attention bias. In their industrial context the use of text, bounding box and therefore not image allows to eliminate an important performance factor in industrial systems. They have conducted several evaluations on several public datasets The Kleister NDA and Kleister Charity, SROIE and CORD. A deep experimental study allows

comparing it to other state-of-the-art methods such as their baseline Roberta, LayoutLM [22], and LayoutLMv2 [21]. Although lambert is a good method, it does not take into account the whole image and the correlation that can exist between the layout information and the textual content. ViBERTgrid[10] proposed a new multi-modal network by combining the best of BERTgrid and CNN to generate a more powerful grid-based document representation. It simultaneously encode the textual, layout and visual information of a document in a 2D feature map.

In order to go one step further, and to reduce the gap between users' need and the documents, we propose to integrate a question answering approach in this kind of architecture. To the best of our knowledge, and as discussed before, different 2D representation of documents have been proposed in the litterature but none of them integrate a question answering in the process and no works have proposed to inclide an attention mechanism mixing the question and the 2D representation.

## 3 Problem Definition

The main objective of our proposal is to provide solutions for the automatic comprehension of documents that requires both a visual analysis and a semantic understanding of their content. The visual analysis remains essential to capture some contextual information. In parallel, the document layout is necessary to extract the correct word in the document (like when a human distinguish two similar tokens based on their visual context). The other hand, automatic comprehension means being able to provide an answer to a question about the content of the document. Starting from a visual context with $D$ documents, we can define the document image $I = i_1, i_2, .., i_d$, its associated bounding boxes (one for each word) $B = b_1, b_2, .., b_m$ and its textual information based on the extracted text (using an OCR) T $= t_1, t_2, .., t_m$ with $m$ being the number of words from the document. We also propose to define the query sentence (with $k$ words) Q=$q_1, q_2, ..; q_k$ which then correspond to the resquest that a user could submit to the system (what he/she is looking for). The questions are related to the corporate document field (i.e. What is the total amount of this invoice? What is this type of document? ) and the answer varies according to the question they can be categorical, numerical or textual.

$$I \in R^d, B \in R^{d \times m}, T \in R^{d \times m} and Q \in R^{d \times k} \tag{1}$$

The questions refer to the document's content. The answers are therefore information extracted from the text present in the document. The objective of our proposed prediction task is then to predict the present beginning and ending word of the answer present in the text. The general assumption is that an unknown function correlates the samples of the questions with the prediction of the beginning word $s1$ and the end word $s2$ , i.e. $(s1, s2) = f(Q)$. The goal of the learning process is to provide an approximation of this unknown function. To better approximate this function $f$, we need to know the correlation between

the question and the answer from the visual and textual features and from the similarity between the questions.

## 4 Proposed approach

In order to set up our proposed model, we got inspired by state-of-the-art models such as Qanet[23] and LayoutLM[22] to propose an architecture based on textual and visual attention models. As we tend to address industrial tasks, where the time needed to process each document must remain low, we only use convolutional and self-attention mechanisms, discarding recurrent neural network (known as slower architectures as they can not process the input tokens in a parallel way).

Another limitation from the most recent and relevant state-of-the-art models relies on their large number of parameters (i.e. 300M parameters for LayoutLMV2). Setting up a model mixing visual content, the semantic, the layout and the question would then require more than 600 million data to converge. We would like to remember that in industrial context, no large annotated dataset could be available, and they would then require a huge effort of annotation to link question and content. We then propose to develop a model able to learn with few data (e.g. hundreds or thousands of annotated documents) while maintaining the execution time as low as possible (generally, companies may not devote more than a second to each document).

Several studies have been conducted to know if we could establish a correspondance between the number of training data and the number of model parameters. The paper [24] conducted several tests on hypothesis related to the generalization capacity of neural networks. The authors demonstrated that "Theoretically, a simple two-layer neural network with $2n + d$ parameters is capable of perfectly fitting any dataset of n samples of dimension d". We can then observe that state-of-the-art approaches have much more than $2n+d$ parameters and one of our hypothesis is to propose a model with much less parameters (less than 10 Millions). In practice, finding the optimal number of parameters is not an easy matter. It depends on the diversity and number of data available in the training set and on the selected architecture. Models with many parameters are likely to overfit while they have the advantage of being able to model much more complicated knowledge (like some latent relationship between the data). Methods with a few numbers of parameters (i.e. less than 20M) are not much studied in the state-of-the-art especially applied on the corporate document. Our second contribution is then to propose a model having a maximum number of 8 millions parameters and which having almost similar results to the state-of-the-art method thanks to our proposed attention mechanisms.
In a first step, we will present the global architecture. Then we will describe the used encoder architecture, and finally we will discuss the proposed co-attention.

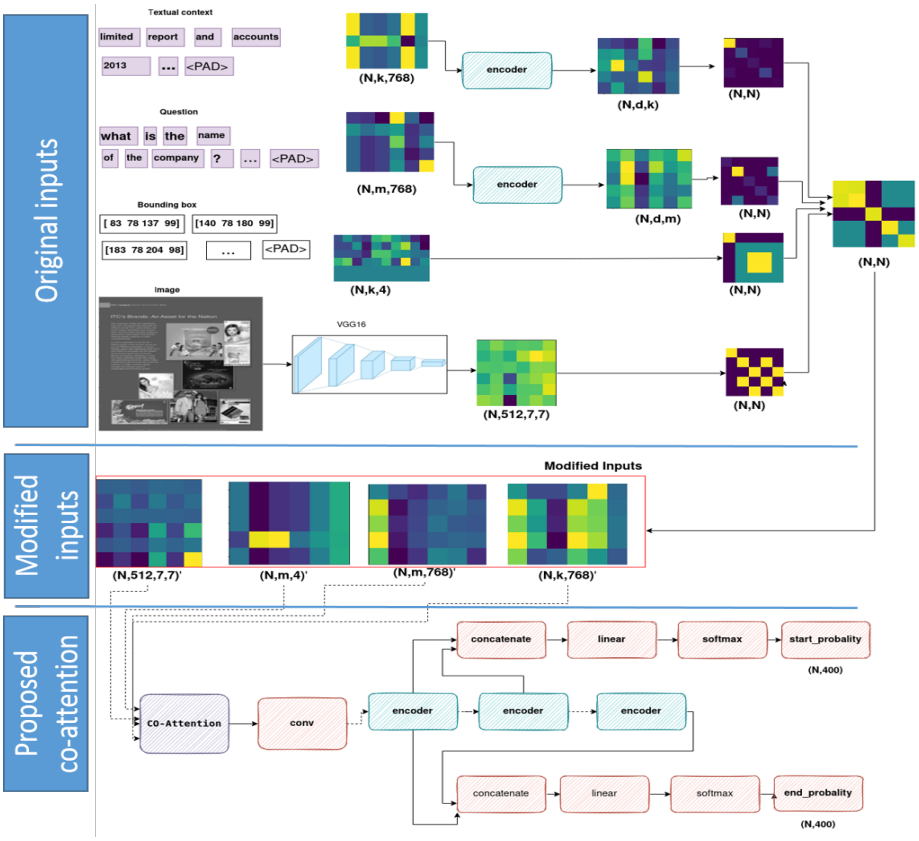


**Fig. 1.** Description of the QAlayout model. The inputs are the textual context extracted from the image, the visual context is whole image . Also the layout information (i.e bounding box of each word) and finally the question. We will try to predict two probality the beginning and end of the answer from the word present in the textual content.We have a self-attention and co-attention mechanism to learn multimodal comprehension

### 4.1 Global description

The global architecture, presented in figure.1, describes the inputs, the intermediate layers and the prediction layer. This last step predicts the probality of having the beginning and the end of the answer. The model has different input features (denoted $T$ , $Q$ and $B$ and $I$ in the equation 2).

The features of each sample are decomposed as follows: $T \in R^{D \times m \times d_1}$ for the textual context , $B \in R^{D \times m \times d_2}$ is the dimension of bounding box for each word of the textual content ,$Q$ is the question input with dimension $R^{D \times k \times d_1}$ and $I$ is the image of the document with the dimension $R^{D \times d_3}$. The dimensions $d_1$ , $d_2$ and $d_3$ are the repective feature dimensions for the textual content, the

bounding boxes and the visual parts. $D$ is the document size of the dataset (i.e $1 \leq i \leq D$).

The probality of the beginning and the end of the answers corresponding the label as $y_i \in R^{2\times m}$ .The beginning and the end correspond to the first and last words that the answer will contain, which is a selection from the words present in the document. The symbole m corresponds the total number of word in the textual context.

The textual content of a document is different, we can have a document with ten words while others exceed 500 words. So that our entries are of the same dimension we keep the prepocessing method described in Qanet to keep one size for all entries. For the semantic embedding of each document, we chose to keep the 400 first words. If the number of word is lower than this, we add some padding tags. We justify this choice of 400 words as we observed that the the average numbers of words in corporate documents is around 300. This parameters could be obviously modified to other context, but we use this value even for larger datasets like DocVQA. In a similar way, we chose to limite the question size to the 50 first words. Once the text has been extracted, we use the classical BERT embedding features [5], where each word is represented by 768 ($d_1$) values. We then obtain a total size of (400,768) for the textual context, and a total size of (50,768) for questions.

For the bounding boxes, we used the outlines of the word from the original image and we normalized them to the width and height of the image. We then obtained a two dimensions vector $d_2$ composed of their cartesian coordinates ($x_{max}$,$y_{max}$,$x_{min}$,$y_{min}$).

The last part of the documents input relies on the visual content. To this end, we resized document images to a dimension vector $d_3 = (224, 224)$. We then use a VGG16 convolutionnel network to embbed the visual content to obtain a (512,7,7) feature dimension vector. This vector is then provided as an input to our co-attention model.

$$\begin{aligned} I &= [i_1, i_2, ..., i_n] \in \mathbf{R}^{D\times d_3} \\ T &= [[t_1, t_2, ..., t_m], ..] \in \mathbf{R}^{D\times m\times d_1} \\ B &= [[b_1, b_2, ..., b_m], ..] \in \mathbf{R}^{D\times m\times d_2} \\ Q &= [[q_1, q_2, ..., q_k], ..] \in \mathbf{R}^{D\times k\times d_1} \end{aligned} \tag{2}$$

As described in the figure 1 ,it's following end-to-end operations:

- In a first step, the input described in the equation goes through models such as Bert and Vgg16 to extract the relevant embedding for our model.
- Then all the input text tensors are passed to the embedding encoding layer which is a single encoding block with 4 conv layers. 8 attention heads are used in the auto-attention module which is the same for all encoding blocks of the model.
- Finally we use self-attention to transform our input into new input, allowing us to exploit the correlation between question and context (e.g. textual context, bounding box and image) at the initial stage.

The self-attention mechanism used in this paper is inspired by [3]. They propose a self-attention mechanism on the input image to consider the inherent correlation (attention) between the input features themselves, and then use a graph neural network for the classification task.

Self-attention allows us to transform $T, B, Q, I$ into $T', B', Q', I'$ which will be the inputs to our co-attention model. To do that, we follow several steps, the first one consists in computing the correlation matrices between the sample and the label.

$$\begin{aligned} C^t &= softmax(TT^T) \\ C^i &= softmax(II^T) \\ C^b &= softmax(BB^T) \\ C^q &= softmax(QQ^T) \end{aligned} \tag{3}$$

Here softmax$(\cdot)$ denotes a softmax operator.The inputs of this function have all the same dimension where $BB^T, QQ^T, II^T and TT^T \in R^{N\times N}$

The self-attention module exploits $C^t, C^i, C^b, C^q$ (see eq 3).

$$C^m = fusion([C^t, C^i, C^b, C^q]) \in R^{N\times N} \tag{4}$$

where $[C^t, C^i, C^b, C^q]$ denotes the attention map concatenation. This fusion function $fusion$ is equivalent e.g $C^m = w_1C^t + w_2C^i + w_3C^b + w_4C^q$, where the weighted parameters $w_1$, $w_2$,$w_3$ and $w_4$ are learned adaptively and N is batch size .

$$\begin{aligned} T' &= TC^m \\ I' &= IC^m \\ B' &= BC^m \\ Q' &= QC^m \end{aligned} \tag{5}$$

These modified inputs as indicated in the equation 5 will be reused in the co-attention part of the model. In the following sections we will detail the encoder part and the co-attention part.

### 4.2 Encoder

As presented in figure 2, the encoder is composed of several convolution layers and a attention layer. Unlike the traditional convolution layer, we use depthwise separable convolution [4]. In this paper [4] depthwise describes that this convolution layer is memory efficient and has better generalization. That help us because this model will be used in a production system and it should be light and fast. For this work, we chose to set up the kernel size to 7, the number of filters to d = 128 and the number of convolutionnal layers within a block to 3. The output of this convolutionnal model is then transfered to an attention layer.

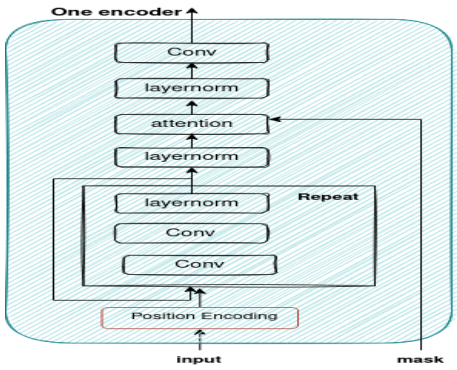


**Fig. 2.** One Encoder

We adopt the multi-headed attention mechanism defined in [17] which, for each position in the input, called a query, computes a weighted sum of all positions, or keys, in the input based on the similarity between the query and the key, as measured by the system. the scalar product.

As one can see in the figure 2, the model involves residual connections, layer normalizations and dropouts too. For each input $x$ and a given operation $f$, the $f$ is defined as $f = (layernorm(x)) + x$ [1] of identity at the input and output of each block repeat. This block is repeated 7 times.

### 4.3 Co-Attention

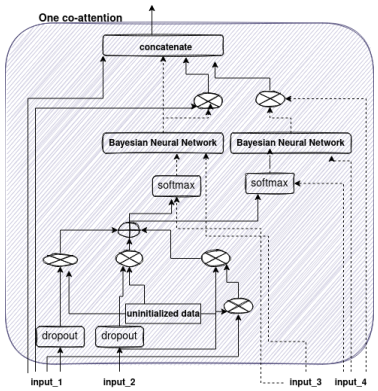


**Fig. 3.** Description of the one co-attention

The proposed "Co-attention" step proposed in our work is inspired by the attention flow layer from the BIDAF architecture [15] (see Fig.3. It calculates

attention in two directions. Context-query attention tells us what query words are the most relevant to each context word like descripted in eq6 and 7. The input_1 ($i_1$) represents the textual context, the boundind box of each word is input_3($i_3$) and input_2 ($i_2$) represents the question finally input_4 ($i_4$)the whole box and encoded context and query respectively. Given that the context length is $j$ and query length is $m$, a similarity matrix is calculated first. The similarity matrix captures the similarity between each pair of context and query words. It is denoted by $S$ and is a n-by-m matrix. The similarity matrix is calculated as,

$$S = f(i_1, i_2) \tag{6}$$

where f is a trilinear similarity function defined as,

$$f(i_1, i_2, i_3, i_4) = W_0[i_1; i_2; i_1 \cdot i_2, i_3 \cdot i_4] \tag{7}$$

## 5 Experiments

### 5.1 Dataset

**Proposed dataset** VQA-CD is a new public dataset containing 3000 questions extracted $\sim$ 693 documents from RVL-CDIP[9]. To the best of our knowledge, no public visual question answering dataset exists for corporate documents. We then decided to annotate and to share our work with the scientific community. This dataset is based on the public RVL-CDIP [9] dataset. This document extracted from RVL-CDIP[9] invoice class does not only contain invoices but also other types of document such as purchase order.

The documents found in VQA-CD dataset contains some documents are well structured while others contain some semi-structured or raw text. This heterogenity of content mimics the industrial context.

When we annotated, we were careful to make sure that each question could be repeated on other documents in order to have a balanced corpus. This balanced corpus is necessary because the inherent understanding of language forces the model to focus on the question rather than the image. For example, if the question is "How much is the ..." and the prediction of the model is a number even if this number does not correspond to the expected one, the model assigns a high probability to it, the same for other type of question ("who is ...", "is it ...?", etc) If for example the question is "What is the total amount" and this question is found in two images I and I'. If the answer related to two images differs then the model will be forced to learn the visual feature to distinguish it.

In order to ensure a kind of compatibility with the DocVQA dataset, we use the same organization of the dataset. We separated our questions for each document randomly. We then divided our 3000 questions into train, test and validation. We took 50% for train and 25% for validation and test.

**SQuAD dataset** Stanford Question [14] Answering Dataset is a reading comprehension dataset consisting of questions posed by crowdworkers on a set of Wikipedia articles. The answer to each question is a segment of text.

This dataset consists of 100,000 questions. The corpus content is only text and no layout or image information is available.

**DocVQA dataset** is to our knowledge the most complete dataset in both content and number of samples. The dataset consists of 50,000 questions defined on 12,000 document images. In the figure 4 you will notice the most recurrent words in the questions and also in the answers.So we can see our questions are often linked to a date or an amount inside in the document.

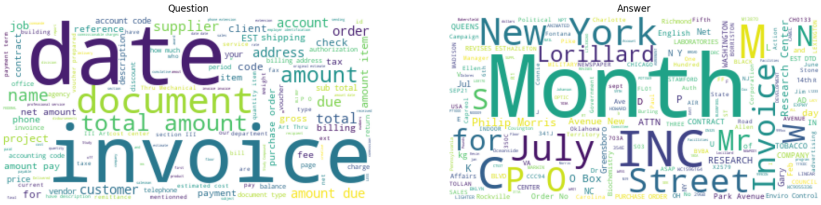


**Fig. 4.** Word clouds of words in answers (left) and questions (right).

**Table 1.** This table contains the results of the proposed QAlayout model and the results of the state-of-the-art method (LayoutLm,Bert).

| Modality | Data | | SQUAD | DOCVQA |
|---|---|---|---|---|
| Method | | Param | F1-SCORE | ANLS |
| Text only | *Bert* | ∼ 110M | 74.430% | 45.57% |
| | **QAlayout(Only_Text)** | ∼ 1M | 82.19% | 48.63% |
| Text + Layout + Image | *LayoutLMv2* | ∼ 426M | — | 86.72 |
| | *LayoutLM* | ∼ 160M | — | 68.93 |
| | **QAlayout(All_inputs)** | ∼ 8M | — | 77.65% |

**Implementation details** The training phase of our model was run with the ADAM optimizer with $\beta 1 = 0.8, \beta 2 = 0.999$ and a batch size of 64, an initial learning rate of 10−3 scaled from 0.1 every 3 epochs without improvement in validation loss and an early stopping after 5 epochs without improvement.

### 5.2 Performance evaluation

To evaluate the performance of our models we will use three metrics (ANLS,F1-SCORE,EM). Average normalized Levenshtein (anls) measure the distance between two string $(q_k, p_k)$ see the equation 8. where Q is the total number of questions. Each question can have 3 answers and k words , $q_k$ the ground truth

**Table 2.** The result for the different categories of question-answer in the Dovqa

| Method | Figure/Diagram | Form | Table/List | Layout | Free_text |
|---|---|---|---|---|---|
| *Bert* | 22.33% | 52.59% | 26.33% | 51.13% | 77.75% |
| *QAlayout(Only_Text)* | 18.53% | 56.12% | 35.85% | 53.42% | 74.75% |
| *QAlayout(All_inputs)* | 39.20% | 73.21% | 86.21% | 63.20% | 71.96% |
| Method | Image/Photo | Handwritten | Yes/No | Others | |
| *Bert* | 48.59% | 35.65% | 3.45% | 5.778% | |
| *QAlayout(Only_Text)* | 30.10% | 37.26% | 17.24% | 34.50% | |
| *QAlayout(All_inputs)* | 44.66% | 62.82% | 59.07% | 67.49% | |

answers and $p_k$ is the prediction of model. $NL(q_k, p_k)$ is the Normalized Levenshtein distance between ground truth and the prediction . Then a threshold $\tau = 0.5$ to filter NL values larger than $\tau$ by returning a score of 0.

$$ANLS = \max_{1..3} s(q_k, p_k)$$
$$s(q_k, p_k) = \begin{cases} (1 - NL(q_k, p_k) & \textbf{if} \quad NL(q_k, p_k) < \tau \\ 0 & \textbf{if} \quad NL(q_k, p_k) >= \tau \end{cases} \tag{8}$$

$$precision = \frac{1 * same\_word}{tail(p_k)} \quad recall = \frac{1 * same\_word}{tail(q_k)} \quad f1 = \frac{2 * precision * recall}{(precision + recall)} \tag{9}$$

The metric F1-SCORE is descripted in 9 . Where same_word count the number of similar words between GT and the prediction.

In the table 1, we have the results on the corpus squad containing only text and the corpus Docvqa containing image and text.

First we tried to compare the performance of the text-only part using the context text and the question only (QAlayout(Only_Text)) with Bert [5].

When we compare our QAlayout(Only_Text) to the state-of-the-art Bert [5] model on the squad corpus we get better results as described in the section. These good performances also add up to a much faster training time. The different attention mechanisms that we have detailed in section Global description have largely contributed to these results. QAlayout(Only_Text) despite its good performance has limitations. These limitations in a text-only corpus may be due to not understanding the syntactic structure. For example if the question is

**Table 3.** The results VQA-CD dataset with different metrics

| Method | ANLS | F1-SCORE | Exact |
|---|---|---|---|
| *QAlayout(Only_Text)* | 36.29% | 29.16% | 25.58% |
| *QAlayout(All_inputs)* | 42.54 % | 35.92% | 33.01% |

*"What is the name of the Bungie Inc. founder who is also a university graduate?"* and the context contains the following words: *"In the arts and entertainment, minimalist composer Philip Glass, dancer, choreographer and leader in the field of dance anthropology Katherine Dunham, Bungie founder and developer of the Halo video game series Alex Seropian, ..."* our model predict "Katherine Dunham" while the correct prediction is "Alex Seropian". Although Katherine is close to the word founder this does not grant her the status of founder. QAlayout(Only_Text) also has other limitations related to the question related to the document layout. For example in the result table, we notice that the QAlayout(Only_Text) is good when the question is related to free_Text and has trouble with the question related to Figure or Form.

The QAlayout method using the image and the characteristic bounding box in addition to the QAlayout(Only_Text) is that allowed to provide answers to the question related to visual structure of the text. The performance has greatly improved on the question related to figure/diagram or form.

In the table 3 we also get the performance of our model on the VQA-CD dataset.

In our VQA for corporate documents, we have either $\sim 80\%$ of the answers with one word and $\sim 17\%$ of contain only two words. Unlike to the other VQA task where the answer size is longer. These one-word answers will certainly result in a kiss on the F1-SCORE. You will notice that the multimodal model is better than the QAlayout(Only_Text) model 5% in all scores . For the VQA-CD corpus we have managed to compute several scores because we have the GT.

The metrics ANLS, F1-SCORE and EM have a value of 100% each if the correspondance between the prediction and GT are totally similar. Their difference comes if the prediction is different from the GT in this case for the em score will have 0% (i.e. either we have all or nothing for this score). For example if the model predicts 1200 and the GT is 100 for the metric ANLS we would have **1 - transformation cost** so NL is 1-0.25=0.75 and as the threshold is equal to 0.5 (i.e the same as the paper [2] ) this result it will be taken into account in the final results. In a document this two numbers (1200,100) can be two different amounts or just a case where the ocr can't extract the character 2 in the image. This score ANLS remains an approximate value that can help in some cases to limit the impact of ocr errors. Finally for the metric F1-SCORE as it is based on the token, we have calculated two tokens one at the word level and another at the character level (i.e. in table 3 it is the token at the word level). The F1-SCORE that we obtained at character level is $\sim 60\%$. It remains clearly higher than the token words. Nevertheless this F1-SCORE metric is not adapted in our task VQA for corporate document because either we take a token at word level and as the majority of our tokens are one word we end up with a very low score or we take at character level and therefore the ANLS score would be more accurate.

Although our model has a good performance, in most cases it confuses the GT amount with another amount or extracts only a part of this answer. The other errors are often OCR errors as the images VQA-CD are old with low-resolution

Also the limitations of QAlayout are also numerous. Sometimes, these limitations are due to a bad understanding of the visual part (i.e difficulty to correlate the elements inside the image). Also the multimodality help us on some cases its performance is not yet the desired one.

## 6 Conclusion and future work

Visual question answering is a task that requires a good understanding of both visual and textual information by correlating this information with the question. We propose a fast and accurate end-to-end method QAlayout that uses visual information the whole image document or layout as well as textual information. Our QAlayout method uses an attention mechanism to take into account the inherent correlation between the question and its visual or textual context at the input of the model with self-attention or after with co-attention. Compared to some state-of-the-art models we have much better results while having less parameters (8M). The limit number of parameters corresponds to the expectation of the industrial context which requires a fast training and prediction time while keeping a reasonable performance. We also contributed to annotate a new dataset VQA-CD containing 3000 questions on corporate documents. Despite the good performance of our model, some limitations exist and we will try to provide a solution. For example, we will build a graph system to establish links between words in the textual content or between areas in the visual content.

## Acknowledgment

This research has been funded by the LabCom IDEAS under the grand number ANR-18-LCV3-0008 , by the French ANRT agency (CIFRE program) and by the YOOZ company.

## References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization (2016)
2. Biten, A.F., Tito, R., Mafla, A., Gomez, L., Rusiñol, M., Mathew, M., Jawahar, C.V., Valveny, E., Karatzas, D.: Icdar 2019 competition on scene text visual question answering (2019)
3. Cheng, H., Zhou, J.T., Tay, W.P., Wen, B.: Attentive graph neural networks for few-shot learning (2020)
4. Chollet, F.: Xception: Deep learning with depthwise separable convolutions (2017)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019)
6. Garncarek, L., Powalski, R., Stanislawek, T., Topolski, B., Halama, P., Turski, M., Gralinski, F.: Lambert: Layout-aware language modeling for information extraction. Lecture Notes in Computer Science p. 532–547 (2021). https://doi.org/10.1007/978−3−030−86549−8$_3$4, http://dx.doi.org/10.1007/978-3-030-86549-8_34

7. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
8. Harley, A.W., Ufkes, A., Derpanis, K.G.: Evaluation of deep convolutional nets for document image classification and retrieval. In: International Conference on Document Analysis and Recognition (ICDAR) (2015)
9. Harley, A.W., Ufkes, A., Derpanis, K.G.: Evaluation of deep convolutional nets for document image classification and retrieval. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR). pp. 991–995. IEEE (2015)
10. Lin, W., Gao, Q., Sun, L., Zhong, Z., Hu, K., Ren, Q., Huo, Q.: Vibertgrid: A jointly trained multi-modal 2d document representation for key information extraction from documents (2021)
11. Lin, Z., Zhang, D., Tac, Q., Shi, D., Haffari, G., Wu, Q., He, M., Ge, Z.: Medical visual question answering: A survey (2021)
12. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach (2019)
13. Mathew, M., Karatzas, D., Jawahar, C.V.: Docvqa: A dataset for vqa on document images (2021)
14. Rajpurkar, P., Jia, R., Liang, P.: Know what you don't know: Unanswerable questions for squad (2018)
15. Seo, M., Kembhavi, A., Farhadi, A., Hajishirzi, H.: Bidirectional attention flow for machine comprehension (2018)
16. TOMSETT, D.: https://digitalhumans.com
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2017)
18. Wang, W., Zhang, J., Li, Q., Hwang, M.Y., Zong, C., Li, Z.: Incremental learning from scratch for task-oriented dialogue systems (2019)
19. Wang, Y., Yao, Q., Kwok, J., Ni, L.M.: Generalizing from a few examples: A survey on few-shot learning (2020)
20. Wu, J., Lu, J., Sabharwal, A., Mottaghi, R.: Multi-modal answer validation for knowledge-based vqa (2021)
21. Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W., Zhang, M., Zhou, L.: Layoutlmv2: Multi-modal pre-training for visually-rich document understanding (2022)
22. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: Layoutlm: Pre-training of text and layout for document image understanding. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Jul 2020). https://doi.org/10.1145/3394486.3403172, http://dx.doi.org/10.1145/3394486.3403172
23. Yu, A.W., Dohan, D., Luong, M.T., Zhao, R., Chen, K., Norouzi, M., Le, Q.V.: Qanet: Combining local convolution with global self-attention for reading comprehension (2018)
24. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning requires rethinking generalization (2017)