ABSTRACT
How far can we go with textual representations for understanding pictures? In image understanding, it is essential to use concise but detailed image representations. Deep visual features extracted by vision models, such as Faster R-CNN, are prevailing used in multiple tasks, and especially in visual question answering (VQA). However, conventional deep visual features may struggle to convey all the details in an image as we humans do. Meanwhile, with recent language models’ progress, descriptive text may be an alternative to this problem. This paper delves into the effectiveness of textual representations for image understanding in the specific context of VQA. We propose to take description-question pairs as input, instead of deep visual features, and fed them into a language-only Transformer model, simplifying the process and the computational cost. We also experiment with data augmentation techniques to increase the diversity in the training set and avoid learning statistical bias. Extensive evaluations have shown that textual representations require only about a hundred words to compete with deep visual features on both VQA 2.0 and VQA-CP v2.

CCS CONCEPTS
• Computing methodologies → Computer vision; Natural language processing; Image representations.

1 INTRODUCTION
The English proverb A picture is worth a thousand words is used to convey the idea that, for humans, the meaning of a complex verbal description is often better understood with a single static image. However, identifying, understanding, and reasoning over the concepts contained in an image is still highly challenging for computer vision systems. Whereas standard deep visual features based on vectors may present some limitations to capture the rich semantic content from a picture [2, 3, 17], descriptive representations based on text may be a good alternative, especially with the recent emergence of powerful language models. This paper aims to provide a deep analysis of textual representations for image understanding tasks, and compare them with existing deep visual features techniques.

A practical task to evaluate image understanding is visual question answering (VQA). VQA aims to answer questions about an image’s visual content, requiring a machine to understand both the question and the image. For describing the visual content, the way in which images are represented is essential, as it highly impacts

![Diagram of the Language-Only Transformer Model](image-url)
the performance of the downstream task. Due to the bottom-up attention’s success [4, 23, 24], deep visual features extracted by object recognition models [48] have been used as the de-facto standard for representing visual content.

On the other hand, the recent progress of Transformer language models [12, 31, 57, 60] produced outstanding advances in several natural language processing (NLP) tasks, e.g., question answering [46] or sentiment analysis [34]. These language models can simultaneously represent inter-relationships between two consecutive sentences and intra-relationships between the individual words in a sentence. With the development of Transformer language models, research on vision-and-language tasks [10, 28, 32, 54, 55] has shifted to explore pre-training methods to learn cross-modal representations on image-text pairs. These pre-trained Transformers are fine-tuned on the downstream vision-and-language tasks, outperforming standard methods based on convolutional neural networks. Within this context, one interesting question arises here: according to the popular proverb, humans may need a thousand words to describe an image; but how many words are necessary for Transformer language models?

In this paper, we explore the effectiveness of textual representations of images and explore if they are competitive with current deep visual features. Specifically, we conduct VQA experiments by representing images using text, instead of deep visual features. Thus, we change the input of a VQA model from image-question pairs to image description-question pairs. We use a RoBERTa [31], a state-of-the-art Transformer language-only model, as a our VQA model. Then, the input description-question pairs are jointly fed into the model to predict an answer (Figure 1). Besides, with the success of data augmentation methods on both VQA [8, 15] and NLP [51, 58, 65] tasks, we investigate the use of synthetic samples on language-only representations. As the aim of the study is to explore the viability of language-only representations in VQA, we rely on already annotated descriptions from two standard datasets [9, 45]. Automatically generating the image descriptions, although a necessary future step, is out of the scope of this paper, and would require a whole paper on its own. We will carefully study image description generation in our future research.

Extensive experiments and analysis, including qualitative and quantitative results, have shown the effectiveness of textual representation of images. We have found that well-described descriptions can outperform deep visual features on two standard VQA datasets, VQA 2.0 and VQA-CP. Additionally, we validated that it does not take a thousand words to describe an image, but about a hundred.

The key contributions of our work are summarized as follows.

- Different from previous VQA models, we use text as image representation. Our setup allows us to study how well textual representation works in VQA. We show that textual representations with about a hundred words are competitive with state-of-the-art deep visual features. This observation promotes more interpretable representations of images in that it gives visibility to what humans easily understand.

- To increase the diversity on the training set, we explore multiple data augmentation techniques. We propose single-modality adaptations to novel VQA data augmentation techniques, as well as we apply standard existing techniques on the NLP community. We find that creating synthetic data consistently improves our language-only model performance. Prominently, back translation for questions boosts the performance by a large margin, not only in our language-only setting but also for deep visual feature-based models.

2 RELATED WORK

Image Representations for VQA. In vision and language tasks, visual features have played a key role in leveraging the visual content of images. Most VQA methods [4, 6, 7, 11, 23, 24, 61] use deep visual features extracted by object detectors, such as Faster R-CNN [48]. These deep visual features are a set of pooled convolutional feature vectors, in which each vector represents a prominent image region. Also, some models that use deep visual features extracted by CNN models, such as Resnet [18], can achieve impressive results [20, 22]. However, VQA models using deep visual features tend to learn superficial linguistic correlations [2, 3, 17].

Building on top of the recent progress in language models, some work has also adapted Transformer models to fuse visual and textual information. Recent studies [10, 28, 32, 54, 55] achieve high performance on vision-and-language tasks, such as VQA, by pre-training multi-layer Transformers with image-caption pairs. As a result, these models can learn general cross-modal representations. Even though they leverage captions of images for pre-training, they rely on deep visual features for image representation.

Similarly, image captioning [62, 63] can be seen as a related task to text representation for images. Yet, converting images into text is different from our purpose of using text for representing images on VQA. Recent work [14] has shown that textual representations for video question answering could outperform models based on deep visual features. Based on these results, we present a rigorous analysis on textual representations for VQA.

Data Augmentation for VQA. Recent work [8, 15] proposed data augmentation techniques for VQA to deal with language bias [2, 3, 17]. They generated counterfactual training samples by masking, removing, or changing critical parts of the input. Training with these synthesized samples makes models focus on the crucial parts in an image and question, leading to alleviate the problem. Inspired by these studies, we propose data augmentation techniques for textual representations.

Data Augmentation for NLP. Various techniques of data augmentation have been proposed to improve NLP tasks. A well-known method is back translation [40, 51, 65] which generates new data by translating sentences to another language and back into the original language. Other techniques such as EDA [58] boosted the performance on text classification tasks [13, 19, 29, 41, 53]. In addition to the data augmentation techniques for VQA, we adopt these methods for our language-only representations, which enables us to check if data augmentation for NLP is also effective for VQA.

3 APPROACH

We introduce the text-based model with which we investigate the potential of language-only representations for VQA. An overview is shown in Figure 1. The input, consisting of a question and a detailed description of the image, is encoded through a Transformer language-only model with several self-attention layers. Then, the
output of the Transformer is fed into a classifier to predict an answer.
We additionally propose the use of data augmentation techniques to increase the size and diversity of the training set.

### 3.1 Language-Only Data

The data for our language-only VQA framework consists of: (1) questions and answers from standard VQA datasets, (2) image descriptions representing image content, and (3) synthetic data obtained from data augmentation techniques.

**Questions and Answers.** For the questions and answers, we use VQA-CP [5] and VQA 2.0 [5] datasets. VQA 2.0 contains 1.1M question-answer pairs of about 204K images, whereas VQA-CP consists of 603K question-answer pairs of about 219K images. All images of VQA 2.0 and VQA-CP are from COCO dataset [30]. Whereas VQA 2.0 has been the de-facto standard for natural image VQA, it has also been shown to contain strong statistical biases on its training distribution [2, 3, 17], by which models obtain high accuracy by only using the first few words of a question. VQA-CP addresses this problem by re-organizing the training and validation splits to have a different answer distribution.

**Image Descriptions.** We obtain image descriptions from two different corpus: COCO captions [9] and Localized Narratives [45]. COCO captions contains five short captions for each image in COCO dataset [30], with an average length of 10.5 words per caption. Captions are obtained by asking annotators to describe the important parts of the scene, without mentioning unimportant details. Localized Narratives provides multimodal image annotations for multiple image datasets (COCO, Flickr30k [64], ADE20K [67], and Open Images [26]). The narratives are generated by asking annotators to describe an image out loud while simultaneously hovering the mouse over the region they are describing, representing the entire image, including minor objects as opposed to COCO captions. This results in an average of 42.9 words per narrative. We only use narratives corresponding to COCO images.

**Synthetic Data.** Additionally, we generate synthetic samples using data augmentation techniques to increase the size and diversity of the training data in our language-only setting. We explore multiple techniques, which can be grouped into two main categories: Data Augmentation for VQA (Section 3.2) and Data Augmentation for Language (Section 3.3). In particular, Data Augmentation for VQA (DAV) are techniques based on the latest multimodal data augmentation methods [8, 15] which propose to generate new images and new questions to increase the diversity in the dataset. Since we do not use the images directly, we propose single modality methods that generate synthetic descriptions that help mitigate the effects of language bias. As for Data Augmentation for Language (DAL), we leverage various established methods within the NLP community for improving language-only tasks [40, 51, 58, 65].

### 3.2 Data Augmentation for VQA

We adapt data augmentation techniques for VQA [8, 15] to our language-only input setting. The aim is to generate new samples that force the model to react to essential parts in the input and provide an answer without relying on the language bias. While image data augmentation methods for VQA [1, 8, 15, 56] mask objects or employ GANs [16] to change the scene in an image, which is computationally expensive, we only require word replacement.

Let \( s \) denote a training triplet, i.e., \( s = (q, d, A) \). \( q = \{q_1, \ldots, q_Q\} \) is a sequence input question with \( Q \) tokens and associated with a question type \( t \), \( d = [d_1, \ldots, d_D] \) is a sequence input description with \( D \) tokens, and \( A = \{a_1, \ldots, a_N\} \) is the set of \( N \) ground truth answers, where \( N \) is at most 10 for the VQA 2.0 and VQA-CP datasets and can vary for different samples. We propose four techniques: (1) hypernym and hyponym replacement, (2) color inversion, (3) adversarial replacement, and (4) counterfactual samples. An example of each technique is shown in Figure 2.

**Hyponym and Hyponym Replacement.** Following [15], we use hypernym and hyponym replacement to introduce similar yet semantically distinct mutations into the image descriptions. For a given word, an hypernym covers a wider range of concepts that the original word implies, e.g. food is an hypernym of fruit. On the other hand, an hyponym covers a narrower range of meanings, e.g. apple is an hyponym of fruit.

To generate new samples, let \( A_d \) denote the set of ground truth answers \( a \) such that \( a \) appears in \( d \). We replace \( a \in A_d \) with its hypernym \( h_a(a) \) (or hyponym \( h_a(a) \)) in both \( A \) and \( d \). The new training triplet for hypernym replacement \( s_h \) is:

\[
\begin{align*}
\mathbf{q}_h & = (\mathbf{q}, \mathbf{d}_h, \mathbf{A}_h) \\
\mathbf{A}_h & = A \setminus A_d \cup \{h_a(a)\}_{a \in A_d} \\
\mathbf{d}_h & = [d_1, \ldots, d_L] \text{ with } d_i = h_a(a) \text{ if } d_i = a \text{ for all } a \in A_d
\end{align*}
\]

(1) (2) (3)

and equivalently with \( h_a(a) \) for hyponym replacement. To avoid duplicates, we do not generate \( s_h \) if \( h_a(a) \in A \) (or \( h_a(a) \in A \)). To identify hypernyms and hyponyms, we use WordNet [37].

---

**Figure 2:** Generated synthetic samples using our proposed data augmentation for VQA techniques.
1Color Inversion. For color inversion [15], we substitute a color word in a description with another color word. First, we manually create a set \( C = \{ c_1, \cdots, c_K \} \) of \( K \) color words, and a set \( T_C \) of question types related to colors.\(^1\) For a given training triplet \( s \) and its question type \( t \), if \( t \in T_C \) and a ground truth color answer \( a \in A \cap C \) is found in the input description (i.e. \( d_i = a \) for some \( i \)), we replace the description and the answer with a different random color word \( c \neq a \). The new training triplet \( s_c \) is:

\[
s_c = (q, d_c, A_c) \tag{4}
\]

\[
A_c = A \setminus \{ a \} \cup \{ c \}
\]

\[
d_c = [d_1, \cdots, d_L] \text{ with } d_i = c \text{ if } d_i = a \tag{5}
\]

Adversarial Replacement. For Yes/No samples, i.e. \( s = (q, d, A) \) such that \( \{ \text{yes, no} \} \cup A \neq \emptyset \), we replace object words \( o \in O \) in description \( d \) with adversarial words, where \( O \) is the set of 80 object classes in COCO dataset [30]. Following [15], we define adversarial word, \( w_{adv}(o) \), as the word that is the most similar yet with a different meaning to \( o \). If \( o \) (or its synonyms) are in \( q \), we change the answer from yes to no; otherwise the answer is not changed. Formally, the new training triplet \( s_a \) is:

\[
s_a = (q, d_a, A_a) \tag{6}
\]

\[
A_a = \begin{cases} 
\{ \text{no} \} & \text{if } o \text{ is in } q \\
A & \text{otherwise}
\end{cases}
\]

\[
d_a = [d_1, \cdots, d_L] \text{ with } d_i = w_{adv}(o) \text{ if } d_i = o, \tag{7}
\]

where \( w_{adv}(o) \) is selected as the closest word to \( o \in O \) according to the Euclidean distance between their Glove embeddings [42].

Differently from [15], we do not generate adversarial samples from questions, as changing or masking a word in a question leads to a new answer, which may not be determined automatically, e.g., “How many bins?” \( \rightarrow \) “How many pens?”.

2Counterfactual Samples. Counterfactual samples [8] are modifications of questions or (part of) images that make the original question-answer pairs irrelevant. We generate counterfactual training samples by adapting [8] to language-only description-question pairs. Given the training triplet \( s = (q, d, A) \), we generate counterfactual samples \( s_{cssq} \) and \( s_{cssd} \) from the query and description:

\[
s_{cssq} = (q_{cssq}, d, A_{cssq}) \tag{8}
\]

\[
s_{cssd} = (q, d_{cssd}, A_{cssd}) \tag{9}
\]

To find \( s_{cssq} \), we feed \( q \) and \( d \) into a trained Transformer VQA model on the target dataset, \( M \), and obtain the contribution of each word in \( q \) to the answer set \( A \) with Grad-CAM [50]. The top-\( D \) words with highest contribution are set as a set \( \Omega_q \) of critical words. Two new questions \( q_{cssq}^+ \) and \( q_{cssq}^- \) are generated by masking all the words in \( q \) not contained in \( \Omega_q \), and masking all the words contained in \( \Omega_q \), respectively:

\[
q_{cssq}^+ = [q_1, \cdots, q_L] \quad \text{with } q_i = <\text{mask}> \quad \text{for all } q_i \notin \Omega_q \tag{10}
\]

\[
q_{cssq}^- = [q_1, \cdots, q_L] \quad \text{with } q_i = <\text{mask}> \quad \text{for all } q_i \in \Omega_q \tag{11}
\]

where \( q_{cssq}^- \) is used as this CSS sample’s question \( q_{cssq} \) and in Eq. (12), the first few words that correspond to the question type (e.g., “what color is”) are not masked, as in [8].

\[
\text{To find } A_{cssq}, q_{cssq}^+ \text{ and } d \text{ are again fed into } M \text{ to obtain the score for each candidate answer. The top-J scoring answers are removed from the original set of answers. Specifically, letting } M_f(q_{cssq}^+, d) \text{ be the set of the top-J answers, the new ground truth answers is given by}
\]

\[
A_{cssq} = A \setminus M_f(q_{cssq}^+, d) \tag{14}
\]

For \( s_{cssd} \), the values of \( \Omega_d, d_{cssd}^+, d_{cssd}^-, \) and \( M_N(q, d_{cssd}) \) are found in the same way.

3.3 Data Augmentation for Language

Given that the input to our VQA model is solely based on the language modality, we also explore NLP data augmentation techniques. Among all existing techniques, we adopt three of the most popular and successful ones [25, 43, 44, 58, 59, 65]: (1) EDA, (2) back translation, and (3) word replacement/insertion via contextual word embedding. Each technique is applied to either the description or the question of the input triplet \( s \) to generate new samples:

\[
s_{nlpq} = (q_{nlp}, d, A) \tag{15}
\]

\[
s_{nlpd} = (q, d_{nlp}, A) \tag{16}
\]

where \( q_{nlp} \) and \( d_{nlp} \) are the question and the description, respectively, after applying one of the transformations below. Examples of synthetic samples are shown in Table 1.

| Original | Synthetic Question |
|----------|--------------------|
| EDA (SR) | What is the camelopard standing behind? |
| EDA (RI) | What is the giraffe abide standing behind? |
| EDA (RS) | What is the standing giraffe behind? |
| EDA (RD) | What is the standing behind? |
| BT       | What’s behind the giraffe? |
| CWR      | What is the giraffe tree behind? |
| CWI      | What is the giraffe standing silently behind? |

| Table 1: Examples of Data Augmentation for Language. |

EDA (Easy Data Augmentation) [58] is composed of 4 operations:

- **Synonym Replacement** randomly chooses \( n \) words from the input sentence and replaces them with their synonyms.
- **Random Insertion** finds a random synonym of a random word in the sentence and inserts it in an arbitrary position.
- **Random Swap** chooses two words and swaps them.
- **Random Deletion** removes words with probability \( p \).

While EDA consists of simple word replacement-based techniques, it has been shown to improve text classification performance in low-resource tasks [58]. For each sample, we apply one of these four operations randomly.

Back Translation [51, 65] translates a sentence into another language and then translates it back into the original language. As shown in [59, 65], back translation can generate diverse paraphrases while preserving the original sentences’ semantics, resulting in significant performance improvements in various NLP tasks [35, 46, 66]. For the implementation of back translation, we use the python nlpqa library [33] based on fairseq [39, 40]. The translator is based on the big Transformer architecture [57] and translates a
We conducted five main experiments to validate the effectiveness of contextual word embedding. To obtain the words that match the context, it uses deep bidirectional language models [12, 43, 60]. We again use the implementation of contextual word embedding provided by python nlpaug library [33]. We replace or insert the most similar words of randomly selected words in the description or question to generate a new one, where we use pre-trained XLNET [60] in the library to obtain the contextual word embeddings.

3.4 Language-Only VQA Model

As opposed to most VQA models, which take image-question pairs as input, the input of our language-only VQA model is description-question pairs. Given the original VQA triplet \((q, I, A)\), where \(I\) is the image associated with the question, we first create the image description \(d\) by concatenating the narrative of Localized Narratives and the captions of COCO captions. The question and the image description are compiled into a single sequence, \(I\), by inserting a classifier token \(<s>\) at the beginning of the sequence, and end of sentence tokens \(</s>\) as:

\[
I = <s> + q + </s> + <s> + d + </s>
\]

where \(\ast\) indicates concatenation. The input sequence is fed into our Transformer-based language model \(T\) to obtain sequence \(f\) of embeddings, i.e.,

\[
f = T(I).
\]

Then, the embedding corresponding to the classifier token \(<s>\) is fed into classifier \(C\), to make a prediction:

\[
p = C(f_{<s>}).
\]

Unless otherwise stated, we use large RoBERTa [31] as our Transformer language model, which has been shown to be one of the best performing models on NLP tasks [38, 49].

4 EXPERIMENTS

We conducted five main experiments to validate the effectiveness of textual representations for VQA: (1) comparison between different image descriptions, (2) comparison between data augmentation techniques, (3) comparing textual representations with deep visual features, (4) evaluating the impact of data augmentation techniques on deep visual feature-based models, and (5) comparison between language-only Transformer.

For RoBERTa hyperparameters, we followed [31], only changing the number of training epochs to 20. As a classifier, we used a multi-layer perceptron with two fully-connected layers with 2048 hidden units, and Swish activation function [47] between them. We use softmax cross entropy over the answer vocabulary for the loss function. The parameters for the data augmentation techniques were as follows: for counterfactual samples, we set \(\text{top-D}=10\) to select critical words and \(\text{top-J}=5\) for assigning new answers; for EDA, we set the rate of the words to be changed to \(0.5\), following the recommended value in the original paper [58]. Unless otherwise stated, the input of our model consists of the whole sequence of question, narrative, and five captions.

Results are presented in terms of accuracy, following the standard protocol [5]. For the datasets split, the VQA-CP v2 training set contains 438K questions about 120K images, whereas the test set has 220K questions about 98K images. As for VQA 2.0, the training set includes 443K questions about 82K images, while the val set contains 214K questions about 40K images. As the VQA 2.0 test set requires access to external server, we report results on the val set.

4.1 Describing an Image with Words

Here, we investigated the relationship between the quality of the image descriptions and the accuracy on the VQA task by conducting two experiments: (1) comparing the performance of different types of image descriptions and (2) studying the correlation between accuracy and the length of the input descriptions.

Image descriptions. We first evaluate the performance of different language-only inputs. Specifically, we consider the following inputs: the question, question and 1 to 5 random captions, question and narrative, and the whole input (question, narrative, and 5 captions). Results on the VQA-CP v2 test set are reported in Table 2, along with the average sequence length.

The whole input, consisting on merging the narrative with the 5 captions with an average of 95.3 tokens per sample, performs the best, which indicates that the two datasets contain complimentary useful information for VQA. When comparing captions and narratives, we find that the former produce better results with fewer

| Image Description | Length | Accuracy |
|-------------------|--------|----------|
| None (Question-Only) | -      | 21.59    |
| 1 Caption         | 10.5   | 35.31    |
| 2 Captions        | 21.0   | 38.49    |
| 3 Captions        | 31.5   | 40.09    |
| 4 Captions        | 42.0   | 41.93    |
| 5 Captions        | 52.5   | 42.34    |
| Narrative         | 42.9   | 36.45    |
| Whole (Narrative + 5 Captions) | 95.3 | 43.64 |

Table 2: Image Description Evaluation. Length denotes the mean number of tokens in the image descriptions.

![Figure 3: Effect of truncating the input sequence on the VQA CP v2 test set. Removing words from the input has a negative effect on the accuracy, suggesting that every word in the sequence contributes to the overall accuracy.](image-url)
words. Specifically, by just using two captions, the performance is already better than with narratives, even though the average number of words is about half. This confirms that the VQA dataset contains a substantial amount of questions about the general content of the image, rather than its specific details, as the captions in COCO dataset, in contrast to narratives, focus mostly on the prominent areas in the scene. When using a single caption, the accuracy improves by 13.92% compared to the Question-Only input, showing that a short description of about 10 words already contains enough information to answer correctly one in three questions. **Input length** For studying the correlation between accuracy and length of the input, we progressively truncate the image descriptions at test time and evaluate their performance on a model trained on the full input without truncation (question, narrative, and 5 captions). To truncate the input, we rearrange the sentences randomly, truncate the words at the end of the rearranged sequence, and put the sentences back in their original order to avoid losing contextual information. Questions are not truncated in any case. Results are shown in Figure 3, where it can be seen that the accuracy decreases as the image descriptions are trimmed. In particular, the accuracy decreases linearly up to 60%. For more than 60% of removal rate, the model cannot deal with the information loss and the accuracy decreases rapidly. This tendency suggests that most words on the input have a positive contribution to the overall accuracy.

### 4.2 Use of Synthetic Samples

We evaluate the performance when augmenting the VQA-CP v2 training set with synthetic samples. The input consists on the whole description with the narrative and the five captions.

Results for each of the proposed data augmentation techniques are shown in Table 3. Whereas most techniques are able to increase the accuracy with respect to the baseline (when no data augmentation is used), back translation for questions has by far the best results, with a gain of 7.52 points on overall, and a boost of 17.30 points on Yes/No questions. As back translation is the only data augmentation technique in Table 3 that generates new samples while maintaining the original semantics, its impressive performance points us to the importance of 1) having diversity within the training questions set, whereas at the same time, 2) a correct relationship between the triplet question-description-answer semantics.

Most of the other DAL techniques, except contextual word replacement/insertion for descriptions, also increase the accuracy with respect to the baseline. An interesting observation is that when data augmentation is applied to the questions, the performance consistently improves more than when it is applied to descriptions. As for the DAV techniques, while all of them improve accuracy, hyponym replacement turns out to be the most powerful one, with an overall accuracy gain of 1.62. Surprisingly, using hyponym and hypernym replacement together does not perform better than hyponym replacement by itself, showing that in some cases the combination of synthetic samples obtained from different techniques may be harmful.

### 4.3 Comparison against Deep Visual Features

We compare our whole language-only representations (question, narrative, and five captions) with state-of-the-art VQA models based on deep visual features on both the VQA-CP v2 and the VQA 2.0 datasets. For a fair comparison, we do not include the models developed to mitigate language bias [6, 7, 11], as these modules can be added as a plug-in extension to any other method, including ours. For the same reason, we do not use data augmentation.

Results are reported in Table 4. Our model outperforms most of the baselines that use deep visual features on both VQA-CP v2 and VQA 2.0. For the accuracy on VQA-CP v2, NSM [21] performs slightly better than our language-only model. This result verifies that the textual representations of the images are effective for both cases where the distribution of answers per question is different between the training and test set and when it is not. This also shows the textual representations capture well the image content, such that they are competitive with deep visual features.

### 4.4 Back Translation for Other Models

The results of data augmentation techniques in Section 4.2 showed that applying back translation to questions is very effective to boost accuracy in our language-only setting. As back translation can be easily applicable to any other settings, we explore if its benefits are transferable to standard VQA models. We conduct experiments on BAN [24] and LXMERT [55] by adding synthetic back translated samples (applied to questions) to the training set.

Results are shown in Table 5. We can see that training BAN and LXMERT with synthesized back translated samples improves the performance by a large margin, with gaps of 2.83 and 3.07 points, respectively. All types of questions are benefited from this technique, especially Yes/No with an improvement of about 6.80. This tendency is consistent with the findings in our model, strongly indicating that leveraging additional samples containing questions with the same meaning but expressed in a different way is extremely beneficial. In other words, adding diversity to the questions increases the model’s ability to interpret the questions.

### 4.5 Language Transformers

Finally, we compare the performance of prevailing language-only Transformer models: BERT [12], XLNET [60], and RoBERTa [31], both in their base and large versions. All the models are exposed to the same input, consisting of the question, the narrative, and the five captions.

Results are shown in Table 6. All the models present a similar behavior. XLNET large has the best performance, and its accuracy is higher 0.59% compared to RoBERTa large. However, while the improvement is minor, the computational time of XLNET large is about 2.7 times that of RoBERTa. Considering this, it is reasonable to use RoBERTa large to save the training time while obtaining relatively similar results.

### 5 WHAT DO TEXTUAL REPRESENTATIONS LEARN?

In this section, we analyze the features and advantages of text for representing images compared with deep visual features. To this end, we conducted the following two analysis: (1) exploring the overlap of the predictions between our language-only model with the models using deep visual features, (2) investigating some visual
Table 3: Data augmentation results on the VQA-CP v2 test set. In DAL, D and Q denote when applied to descriptions or questions, respectively. Gap is the overall accuracy difference compared to the accuracy when not using synthetic samples.

| Input Data                                      | Num. Synthetic | Num. Total | Yes/No | Number | Other | Overall | Gap  |
|------------------------------------------------|----------------|------------|--------|--------|-------|---------|------|
| Narrative + 5 Captions                          |                | 438,183    | 45.13  | 20.06  | 49.33 | 43.64   | -    |
| w/ Hyponym Replacement                          | 132,570        | 570,753    | 45.65  | 25.36  | 50.52 | 45.26   | +1.62|
| w/ Hypernym Replacement                         | 23,869         | 462,052    | 47.28  | 17.69  | 49.10 | 43.70   | +0.06|
| w/ Hyponym and Hypernym Replacement             | 183,944        | 622,177    | 45.80  | 21.46  | 51.15 | 45.06   | +1.42|
| w/ Color Inversion                              | 19,308         | 457,491    | 45.61  | 19.93  | 50.60 | 44.47   | +1.06|
| w/ Adversarial Word Replacement                 | 169,929        | 608,112    | 44.71  | 19.84  | 50.63 | 44.48   | +0.29|
| w/ Counterfactual Samples                       | 438,183        | 876,366    | 44.20  | 19.84  | 52.07 | 44.86   | +1.22|
| DAL w/ EDA (D)                                  | 438,183        | 876,366    | 44.68  | 20.64  | 50.08 | 44.02   | +0.38|
| DAL w/ EDA (Q)                                  | 438,183        | 876,366    | 46.86  | 23.50  | 50.62 | 45.39   | +1.75|
| DAL w/ Contextual Word Replacement (D)          | 438,183        | 876,366    | 44.69  | 19.40  | 48.91 | 43.18   | -0.46|
| DAL w/ Contextual Word Replacement (Q)          | 438,183        | 876,366    | 46.09  | 22.49  | 49.10 | 44.16   | +0.52|
| DAL w/ Contextual Word Insertion (D)            | 438,183        | 876,366    | 45.15  | 19.31  | 48.86 | 43.27   | -0.37|
| DAL w/ Contextual Word Insertion (Q)            | 438,183        | 876,366    | 45.86  | 21.44  | 51.10 | 45.05   | +1.41|
| DAL w/ Back Translation (D)                     | 438,183        | 876,366    | 45.28  | 21.01  | 50.89 | 44.70   | +1.06|
| DAL w/ Back Translation (Q)                     | 293,811        | 731,994    | 62.43  | 27.15  | 51.84 | 51.16   | +7.52|

Table 4: Comparison of language-only representations with standard deep visual features. * indicates our re-implementations.

| Model               | VQA-CP v2 test | VQA 2.0 val |
|---------------------|----------------|-------------|
|                     | Yes/No | Number   | Other | Overall | Yes/No | Number | Other | Overall |
| HAN [36]            | 52.25  | 13.79    | 20.33 | 28.65   | -      | -      | -     | -       |
| RAMEN [52]          | -      | -        | -     | -       | 39.21  | -      | -     | -       |
| MuRel [6]           | 42.85  | 13.17    | 45.04 | 39.54   | -      | -      | -     | 65.14   |
| UpDn [4]            | 42.27  | 11.93    | 46.05 | 39.74   | 81.18  | 42.14  | 55.66 | 63.48   |
| ReGAT [27]          | -      | -        | -     | 40.42   | -      | -      | -     | 67.18   |
| BAN* [24]           | 43.14  | 13.63    | 46.92 | 40.74   | 83.19  | 48.13  | 57.52 | 65.93   |
| LXMERT* [55]        | 42.01  | 14.16    | 48.34 | 41.28   | 83.30  | 46.15  | 56.91 | 65.31   |
| NSM [21]            | -      | -        | -     | -       | 45.80  | -      | -     | -       |
| Ours (Narrative + 5 Captions)                  | 45.13  | 20.06    | 49.33 | 43.64   | 87.91  | 56.47  | 59.43 | 69.74   |

Table 5: Results of applying back translation to different VQA models in the VQA-CP v2 test set. Gap represents the improvement by training with the synthetic samples.

| Model               | Yes/No | Number   | Other | Overall | Gap |
|---------------------|--------|----------|-------|---------|-----|
| BAN [24]            | 43.14  | 13.63    | 46.92 | 40.74   | -   |
| w/ back translation | 47.87  | 16.27    | 48.76 | 43.57   | +2.83|
| LXMERT [55]         | 42.01  | 14.16    | 48.34 | 41.28   | -   |
| w/ back translation | 48.81  | 16.16    | 49.75 | 44.35   | +3.07|
| Ours                | 45.13  | 20.06    | 49.33 | 43.64   | -   |
| w/ back translation | 62.43  | 27.15    | 51.84 | 51.16   | +7.52|

Table 6: Evaluation of different language-only Transformer models on the VQA-CP v2 test set.

| Model               | Yes/No | Number   | Other | Overall |
|---------------------|--------|----------|-------|---------|
| BERT base           | 42.66  | 16.19    | 45.66 | 40.29   |
| BERT large          | 42.72  | 17.43    | 48.47 | 42.06   |
| XLNET base          | 43.49  | 17.61    | 48.45 | 42.30   |
| XLNET large         | 44.58  | 20.67    | 50.52 | 44.23   |
| RoBERTa base        | 44.39  | 17.46    | 48.74 | 42.70   |
| RoBERTa large       | 45.13  | 20.06    | 49.33 | 43.64   |

5.1 Error Analysis

We compare our model that only takes text as input with models that use deep visual features and investigate whether our model makes the same or different mistakes. We adopt BAN and LXMERT as representative models leveraging deep visual features. Whereas BAN was one of the top-performing models before the emergence of the Transformer-based vision-and-language models, LXMERT is one of the state-of-the-art pre-trained models based on multi-modal Transformers extracting cross-modal features. LXMERT can be seen as the middle model between our language-only model and BAN, as LXMERT is pre-trained by taking a pair of image and caption as input.

Figure 5 compares our model with BAN and LXMERT in terms of the consistency of correct and incorrect answers. Note that the examples for qualitative analysis. The input to our models is formed by combining the narrative with the five captions.
5.2 Qualitative Analysis

We show some qualitative examples in Figure 4. We compare our model’s predictions with the predictions of BAN and LXMERT. Example (1) asks what the man is doing with his hand. The image description contains the expression of what the man is doing (“A man pointing to pots…”), result in our language-only model answering correctly. On the other hand, BAN and LXMERT fail to make the correct prediction, which they answer “cooking” nevertheless the man is not cooking. The object detector can detect the cooking tools, so the models may cause this mistake guess from the utensils, not from the man’s move. In example (2), the image description also describes the critical information to answer the question (“Six snowboards”), while the object detector cannot detect all objects. On the contrary, examples (3) and (4) show the limitations of our language-only model. Example (3) requires reading the letters on the side of the plane. The image description contains the necessary word (“AirFrance”) to answer the question but fails to make a correct prediction. This result poses the lack of language-only ‘Transformer models’ ability to understand the contents and relations between the question and description. Additionally, example (4) shows the importance of the image descriptions’ quality. Our model fails to correctly answer because there is no information to answer the question. From these observations, utilizing well-described text as image representation has the advantage against the deep visual features when answering the questions that deep visual features do not work well. On the other hand, we identify the limitation of our language-only model for understanding the text input.

5.3 Limitations

Lastly, we analyze the limitation of our approach. Note that it is not fair to directly compare the results by our language-only model and the deep visual feature-based models, as our approach uses additional annotations that deep visual features does not, i.e., annotated sentences from Localized Narratives and COCO captions. This means that the sentences describe the content prominent for human eyes. The questions in the dataset, also annotated by humans, are additional annotations that deep visual features does not, i.e., annotated sentences from Localized Narratives and COCO captions. This means that the sentences describe the content prominent for human eyes. The questions in the dataset, also annotated by humans, are based on the same perspective, and our model may take this advantage. Meanwhile, deep visual features are trained in the end-to-end manner, which can provide strong supervision on what to see. This on the other hand benefits deep visual feature-based models. Yet, our results give interesting insights on differences and analogies between deep visual features and textual representation, providing...
a baseline for the VQA tasks with the interpretable representation. Moreover, this study brings us the opportunity to introduce a new research direction for VQA in particular, and image understanding in general: to automatically generate image description as image representations instead of (or combined with) deep visual features.

6 CONCLUSION

In this paper, we explored using textual representations of images instead of deep visual features for VQA. Additionally, we explored data augmentation methods for descriptions and questions to increase the training data size and diversity. Through the experiments, including ablations, we validated our language-only model competed with deep visual feature-based models. Also, most of the data augmentation techniques enhanced the performance; especially the boost by back translation for questions was outstanding. It was revealed that machines do not need to take a thousand words to understand an image, but only a hundred. In this study, we dealt with VQA to validate the effectiveness of the textual representations, but these representations may be potentially applied to other tasks, not only VQA. In future work, we will 1) delve into why back translation for questions works so well; 2) delve into the textual representations for various tasks; 3) explore the performance when using a text generator to make the image descriptions.

REFERENCES

[1] Vedika Agarwal, Rakshith Shetty, and Mario Fritz. 2020. Towards Causal VQA: Re-vealing and Reducing Spurious Correlations by Invariant and Covariant Semantic Editing. In CVPR. IEEE, 9647–9655.

[2] Ashwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the Behavior of Visual Question Answering Models. In EACL. The Association for Computational Linguistics, 1955–1960.

[3] Ashwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don’t Just Assume: Look and Answer: Overcoming Priors for Visual Question Answering. In CVPR. IEEE Computer Society, 4971–4980.

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In CVPR. IEEE Computer Society, 6077–6086.

[5] Stanislaw Antol, Ashwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In ICCV. IEEE Computer Society, 2425–2433.

[6] Rémi Cadène, Corentin Danicette, Hedi Ben-younes, Matthieu Cord, and Nicolas Thome. 2019. MUREL: Multimodal Relational Reasoning for Visual Question Answering. In CVPR. Computer Vision Foundation / IEEE, 1989–1999.

[7] Rémi Cadène, Corentin Danicette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. 2019. RUBiS: Reducing Unimodal Biases for Visual Question Answering, In NeurIPS. IEEE, 839–850.

[8] Long Chen, Xin Yan, Jun Xiao, Huanwang Zhang, Shiliang Pu, and Yueting Zhuang. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. CoRR abs/1504.00325 (2015).

[9] Yen-Chun Chen, Linjie Li, Chitta Baral, and Yezhou Yang. 2020. A Training Paradigm for Out-of-Distribution Generalization in Visual Question Answering. In EMNLP (1). Association for Computational Linguistics, 878–892.

[10] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In NIPS. 2672–2680.

[11] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. MUTANT: A Training Paradigm for Out-of-Distribution Generalization in Visual Question Answering. In EMNLP (1). Association for Computational Linguistics, 878–892.

[12] Jian J. Gouwselloff, Jean Pouget-Abadie, Mehdi Milzra, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In NIPS. 2672–2680.

[13] Noa Garcia and Yuta Nakashima. 2020. Knowledge-Based Video Question Answering with Unsupervised Scene Descriptions. In ECCV (18) (Lecture Notes in Computer Science, Vol. 12363). Springer, 581–598.

[14] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. MUTANT: A Training Paradigm for Out-of-Distribution Generalization in Visual Question Answering. In EMNLP (1). Association for Computational Linguistics, 878–892.

[15] Jian J. Gouwselloff, Jean Pouget-Abadie, Mehdi Milzra, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In NIPS. 2672–2680.
