Neurocomputing 000 (2018) 1–8

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A flexible testing environment for visual question answering with performance evaluation☆

Mihael Cudic*, Ryan Burt, Eder Santana, Jose C. Principe

The University of Florida, Gainesville, FL 32601, United States

ARTICLE INFO

Article history:
Received 30 January 2017
Revised 20 April 2017
Accepted 20 February 2018
Available online xxx

Communicated by Prof. Zidong Wang

Keywords:
Data environment
Deep learning
Visual Question Answering

ABSTRACT

In order to move toward efficient autonomous learning, we must have control over our datasets to test and adaptively train systems for complex problems such as Visual Question Answering (VQA). Thus, we created a testing environment around MNIST images with optional cluttering. Although less complex than publicly available VQA datasets, the new environment generates datasets that decouple answers from questions and incorporate abstract ideas (content, context, and arithmetic) that must be learned. In addition, we analyze the performance of merged CNNs and LSTMs using the environment while exploring different ways to incorporate pretrained object classifiers. We demonstrate the usefulness of our environment as well as provide insight on the limitations of simple architectures and the complexities of different questions.

1. Introduction

With increased processing power and advancements in machine learning, Artificial Intelligence (AI) can solve problems that go beyond the recognition of objects [1–4]. From natural language processing (NLP) [1,5] to game AIs [6], these machines are approaching human capabilities, sometimes surpassing us.

These deep learning architectures have proven to be especially adept at understanding visual scenes [7]. More recently, these architectures have moved beyond simply labeling the main object contained in images to incorporating extra information such as image rankings [8] or poses [9]. In addition, new architectures are separating out multiple objects within scenes, labeling them, and then performing additional tasks such as blurring for privacy [10].

However, the tasks and the performance of these machines are limited to the datasets provided. Ideally, all architectures should be trained on large enough datasets to accurately represent the distribution of samples in the dataspace. For some tasks, such as MNIST [11], this goal is attainable because the number of classes is known and one can collect sufficient data to be representative; but, for tasks that involve an undetermined number of objects or relations, this goal becomes much harder to reach. Therefore, during initial development and to speed up evaluation, there is a need for flexible synthetic environments that are under the control of the developer and can be practically unbounded to test the generalization of the machine learning models.

One of the next steps for these machines is answering questions given an image, or more commonly known as Visual Question Answering (VQA) [12–22]. In order to successfully perform this task, these machines not only have to recognize what is asked, but also understand the context of the objects that it must classify. This requires 3 main skills: object recognition ("What animal is in the picture?), contextual understanding ("Where is the ball?"), and advanced reasoning ("Are there enough bricks to build the house?"). Yet, the only way to evaluate a machine's performance is by looking at its classification error over the entire dataset.

Thus, providing large datasets are not sufficient to train machines to learn problems. If any of the individual relations required by the task are underrepresented in the dataset, then the machine will have trouble learning the task and will not generalize well. To move toward autonomous learning, there needs to be a flexible and controlled synthetic environment where the user can easily generate novel questions that pairs the input and output data with all the necessary information. This will allow proper evaluation of our machine's performance on individual scenarios and adaptation of future datasets to maximize performance on all skills required to complete the task. Despite synthetic datasets generally being simpler than real world datasets, they can be useful for prototyping new architectures and breaking down fault categories.

☆ This research was funded by ONR grant N00014-14-1-0542.

* Corresponding author.
E-mail addresses: mcudic@ufl.edu (M. Cudic), rburt@ufl.edu (R. Burt), edersantana@ufl.edu (E. Santana), principe@cnel.ufl.edu (J.C. Principe).

https://doi.org/10.1016/j.neucom.2018.02.065
0925-2312/

Therefore, in this paper, we propose a synthetic flexible VQA data environment using MNIST images pasted onto a canvas with corresponding decoupled questions. The data was then used to test current baseline architectures and various hyper parameters in order to understand their current limitations. In addition, we test different ways to incorporate pretrained object classifiers to mimic the use of ImageNet classifiers on larger VQA neural network architectures.

2. Related works

2.1. VQA datasets

There are some publicly available VQA datasets made to test various architectures. However, these datasets are not controlled and test a narrow set of predetermined relations, hindering our understanding of the architecture's full performance and bottlenecks.

Some datasets only contain "Yes" or "No" [12] questions while others only contain images of a predefined world with restricted objects [13]. COCO-QA consists of images containing common objects and asks four types of questions pertaining to object, number, location, and color [14]. FM-IQA extends this by also using images from COCO, but image annotations and questions are crowdsourced to introduce more variability in the questions [15]. The Visual Genome provides a large dataset and breaks image information into three main groups: objects, attributes, and the relationships between objects [16]. Visual7w is another dataset that places a large emphasis on relationship between objects and uses language that precisely describes these relationships [17]. Lastly, the Virginia Tech VQA dataset provides a large scope by generating 250,000 real world and synthetic images, with 76,000 questions and 10 million answers [18].

These datasets are also limited in their post-analysis capabilities. Although the Virginia Tech VQA dataset does distinguish questions by phrasing, it does not distinguish the skills required to answer such questions (i.e. object recognition, contextual understanding, or advanced reasoning). More so, the dataset relies predominately on questions that require object classification. Questions such as "What sport…" or "What color…" all require an object as an answer. Even yes or no questions could require object recognition: "Is that a tree?" or "Are they on the beach?".

In addition, there are examples when questions asked require some form of contextual understanding. VQA tasks generally require the network to learn two modes of data: one for understanding the questions and another for extracting data from the environment. However, these questions can often be answered without understanding the context of the picture, using only a single data mode to represent both. By only looking at the answers to previous questions, an LSTM has an accuracy of 48.76% on open ended questions on the Virginia Tech VQA dataset [8], which we think it is too high to be realistic. Instead, we using our MNIST VQA environment, an LSTM has a 31.9% accuracy when only looking at the questions. This means that the questions and answers are more decoupled, requiring architectures to look at the image rather than correlations in questions and answers, effectively learning both modes rather than representing both sets of information with a single mode.

Therefore, there are two main distinctions between our environment and previous VQA datasets: the ability to generate new image/question pairs to fit a specific task and the separation between the question information and the image information. The ability to easily generate new questions and corresponding images gives testers control over the environment that will allow them to test specific features of different architectures. This includes accounting for a skew in the datasets which often occurs in current real world VQA environments. And finally, the separation ensures that any learning architecture will need to extract information from both sources to answer the question rather than learning a shortcut by optimizing the knowledge gained from a single source.

2.2. Architectures

There have been multiple approaches to VQA using CNNs and RNNs. Malinowski, Rohrbach and Fritz feed both the image features produced by a CNN and the corresponding question words into an LSTM to generate word answers [19]. ACVT Adelaide used a multi-label CNN to generate the top 5 attributes of each image (attention) which were then used to create image captions (caption) and brief text documents based on their Wikipedia summaries (knowledge). The attention, caption, knowledge and question components were inputted into an LSTM which outputted word answers [20]. Alternatively, UC Berkeley and Sony used a CNN and LSTM which feed visual and textual embeddings into a multimodal bilinear compact model to output word answers [21].

However, the simplest of architectures includes merging the outputs of a CNN and RNN with additional fully connected layers [18,22]. Regardless of its simplicity, this architecture performed relatively well on the VQA dataset with an accuracy of 57.75% on open ended image questions. By combining an RNN for the question and a CNN for the image, we have two parallel architectures that learn the two data modes separately. Due to its simplicity and relatively good performance, this architecture will be used as the model for our baseline experiments. Newer implementations should outperform this baseline on our MNIST VQA dataset; however, using this baseline will show its limitations and bring forth the architectural improvements needed to perform better on larger VQA datasets.

3. MNIST VQA data environment

We created an environment that outputs a sample dataset of canvases and corresponding questions and answers. The canvases contain randomly pasted MNIST images with the option of cluttering. The questions currently come in three major abstract concepts: content, context, and arithmetic. The answers are the appropriate responses to the question given a specific canvas.

3.1. Canvas

As shown in Fig. 1, the generation of the canvas has multiple parameters. The user can specify a minimum and maximum number of MNIST images pasted as well as a minimum and maximum scaling for each image pasted (Fig. 1a and b). These values are uniformly generated for each new canvas with each pasted MNSIT assigned an independent scale.

The user can also specify the x&y bordering of the canvas (Fig. 1b). A positive border parameter ensures that no MNIST image will enter the border space; thus, limiting the potential places for the image to be pasted. Conversely, a negative border parameter allows for an MNIST image to enter an artificial space outside the canvas; thus, increasing the potential places for an image to be pasted. However, the border parameter does not affect the original canvas dimensions as the canvas will just be cropped to its original dimensions. This is useful for datasets that have a lot of empty space surrounding the important information in the image, such as MNIST.

Additionally, the user can specify the minimum pixel separation between the centers' of each image pasted (Fig. 1c) in order to avoid potential overlap between pasted images. The minimum pixel separation parameter assumes both centers belong to 28×28

Please cite this article as: M. Cudic et al., A flexible testing environment for visual question answering with performance evaluation, Neurocomputing (2018), https://doi.org/10.1016/j.neucom.2018.02.065