An Empirical Evaluation of Visual Question Answering for Novel Objects

Santhosh K. Ramakrishnan[1,2] Ambar Pal[1] Gaurav Sharma[1] Anurag Mittal[2]
[1]IIT Kanpur* [2]IIT Madras†

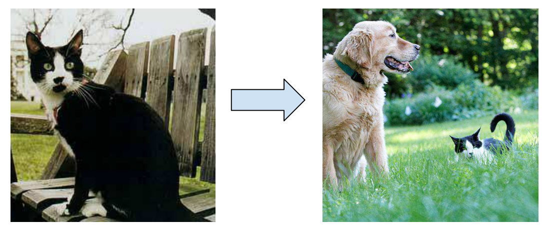
Abstract

We study the problem of answering questions about images in the harder setting, where the test questions and corresponding images contain novel objects, which were not queried about in the training data. Such setting is inevitable in real world—owing to the heavy tailed distribution of the visual categories, there would be some objects which would not be annotated in the train set. We show that the performance of two popular existing methods drop significantly (up to 28%) when evaluated on novel objects cf. known objects. We propose methods which use large existing external corpora of (i) unlabeled text, i.e. books, and (ii) images tagged with classes, to achieve novel object based visual question answering. We do systematic empirical studies, for both an oracle case where the novel objects are known textually, as well as a fully automatic case without any explicit knowledge of the novel objects, but with the minimal assumption that the novel objects are semantically related to the existing objects in training. The proposed methods for novel object based visual question answering are modular and can potentially be used with many visual question answering architectures. We show consistent improvements with the two popular architectures and give qualitative analysis of the cases where the model does well and of those where it fails to bring improvements.

1. Introduction

Humans seamlessly combine multiple modalities of stimulus, e.g. audio, vision, language, touch, smell, to make decisions. Hence, as a next step for artificial intelligence, tasks involving such multiple modalities, in particular language and vision, have attracted substantial attention recently. Visual question answering (VQA), i.e. the task of answering a question about an image, has been recently introduced in a supervised learning setting [21, 3]. In the currently studied setup, like in other supervised learning settings, the objects in the training data and the test data overlap almost completely, i.e. all the objects that appear during testing have been seen annotated in the training. This setting is limited as this requires having training data for all possible objects in the world—this is an impractical requirement owing to the heavy tailed distribution of the visual categories. There are many objects, on the tail of the distribution, which are rare and annotations for them might not be available. While humans are easily able to generalize to novel objects, e.g. make predictions and answer questions about a wolf, when only a cat and/or a dog were seen during training, automatic methods struggle to do so. In the general supervised classification, such a setting has been studied as *zero shot learning* [15], and has been applied for image recognition as well [11, 14, 36, 40]. While the zero shot setup works with the constraint that the test classes or objects were never seen during training, it also assumes some form of auxiliary information to connect the novel test classes with the seen train classes. Such information could be in the form of manually specified attributes [11, 14, 40] or in the form of relations captured between the classes with learnt distributed embeddings like, `Word2Vec` [23] or `GloVe` [25], of the words from an unannotated text corpus [36]. In the present paper, we are interested in a similar setting, but for the more unconstrained and challenging task of answering questions about novel objects present in an image. Such a setting, while being natural, has not been studied so far, to the best of our knowledge.

Train on known objects   Test on unknown objects



*Is the **cat** black and white ?* *Yes*   *Is the **dog** black and white ?* ___

Figure 1: We are interested in answering questions about images containing objects not seen at training.

We start studying the problem by first proposing a novel split (§4.1), into train and test sets, of the large-scale pub-

*The project started when Santhosh Ramakrishnan and Ambar Pal were summer interns at IIT Kanpur. Ambar Pal is a student at IIIT Delhi. `ambar14012@iiitd.ac.in, grv@cse.iitk.ac.in`
†`{ee12b101@ee, amittal@cse}.iitm.ac.in`

lic dataset for VQA recently proposed by Antol et al. [3]. Our split ensures that the novel objects in the test set are never seen in the train set; we select the novel objects and put all the questions that contain those objects into the test set along with all the questions whose answers contain them as well. This means that the train set does not contain any question which either (i) makes a query about the novel objects, or (ii) queries about some aspect of the image which indicates any of the novel objects, i.e. has any possible answer mentioning the novel object. Hence, the split is strong as any information about the novel object is missing from the train set.

We then take two deep neural network based architectures which have shown good performance on tasks based on language and vision combined [18, 26]. We benchmark them on the new split for novel object VQA and compare the performances on the known object setting. As expected, we find that the performances drop significantly (up to 28%) when there are novel objects in the test set. We then propose two methods based on deep recurrent neural network based multimodal autoencoder, which exploit large existing auxiliary datasets of text and images, to answer questions about novel objects, with the two architectures studied. The proposed frameworks are modular and can be used with many neural networks based VQA systems. We show that the proposed methods improve the performance of the system, equally when (i) an oracle is assumed, that gives the novel test objects and (ii) when the minimal assumption is made that the novel test objects are semantically related, quantified by their similarity in distributed `Word2Vec` embedding space [23], to the train objects. We extensively study multiple configurations quantitatively and also analyse the results qualitatively to show the usefulness of the proposed method in this novel setting.

2. Related works

Image based question answering was introduced by Malinowski and Fritz [20] as the *Visual Turing Test*. With the large scale dataset, introduced by Antol et al. [3], recently there has been a lot or interest in the problem. The survey by Wu et al. [34] categorizes the methods for VQA into three categories. First, the joint embedding based methods, which bring the visual and textual vectors into a common space and then predict the answer [8, 9, 21, 24, 27, 1, 2], second, attention based systems which focus on the relevant spatial regions in the images which support the question [5, 19, 29, 38, 39, 41] and finally, third, which are based on networks with explicit memory mechanisms [13, 37].

Malinowski et al. [21] and Gao et al. [9] encode the question and image using an LSTM and use a decoding LSTM to generate the answers. Ren et al. [27] predict a word answer using a multi-class classification over a pre-defined vocabulary of single word answers. Fukui et al. [8] propose a multimodal bilinear pooling, using Fourier space computations for efficiency.

Zhu et al. [41] augment the LSTM with spatial attention, by learning weights over the convolutional features. Similarly, Chen et al. [5] generate a question-guided attention map using convolution with a learnt kernel. Yang et al. [39] use stacked attention networks that iterate to estimate the answer. Xu et al. [38] propose a multi-hop image attention scheme, where the two types of hops are guided by word-based and question-based attention. Shih et al. [29] use region proposals to find relevant regions in the image w.r.t. the question and potential answer pairs. Lu et al. [19] propose a hierarchical co-attention model where both image and question steer the attention over parts of each other.

Dynamic Memory Networks of Kumar et al. [13] and their variants [33, 30, 4], have been recently adapted and applied to VQA by Xiong et al. [37]. They use an explicit memory to read and write depending on the input question, allowing them to understand the questions better.

Methods which use auxilary image or text datasets or other sources of knowledge have also been proposed. Wang et al. [31, 32] propose methods which use knowledge bases for VQA. Wu et al. [35] predict semantic attributes in the image and exploit external knowledge bases to query for related knowledge, to better understand the question.

Similar in spirit to the current work, zero shot learning, i.e. when the set of test classes is disjoint from the set of train classes, has been well studied in the literature [11, 14, 15, 40]. Zero shot learning aims to predict novel object categories without any visual training examples but with auxilary relations between the known and unknown objects, e.g. in the form of common attributes. Lampert et al. [14] proposed to use attributes for zero shot image classification while more recent work by Xian et al. [36] showed that it could be achieved using embeddings learnt from unsupervised text data. Most of the current state-of-the-art methods for zero shot classification use an embedding based approach where the images and classes (the word for the class, e.g. 'dog', 'cat') are embedded into respective spaces and a bilinear compatibility function is learnt to associate them [7, 36].

Our work is also related to the recent works on autoencoders for vector sequences based on recurrent neural networks (RNN). Such autoencoders have been recently used in text processing [16, 6] as well for doing semi-supervised learning and fine tuning of RNN based language models.

3. Approach

We are interested in extending the VQA models to better answer questions about novel objects by being aware of them both textually and visually. Towards that end, we start with two existing architectures, for VQA, and expose them to extra information, from auxiliary datasets of text and images, in a carefully designed manner. This allows them to be able to answer questions about novel objects that are not present in the VQA training data. We consider two successful deep neural network based architectures, illustrated in Figure 2, whose variants have been used in recent literature [18, 26]. We first describe the base architectures and then give the proposed training and architectural extensions for novel object induction.

3.1. Base Architectures

Architecture 1. The first architecture, shown in Figure 2 (left), proposed by Lu et al. [18], uses a Long Short Term Memory (LSTM) based recurrent neural network, to encode the question as $\mathbf{x}_Q \in \mathbb{R}^{d_Q}$, and a Convolutional Neural Network (CNN) to encode the image as $\mathbf{x}_I \in \mathbb{R}^{d_I}$. The two encoded representations are then projected to a common multimodal space with projection matrices $W_Q \in \mathbb{R}^{d\times d_Q}$ and $W_I \in \mathbb{R}^{d\times d_I}$ respectively. The projected vectors are then multiplied element-wise to obtain the joint multimodal representation of the question and the image. This representation is then, in turn, projected to the answer space using a fully connected layer to obtain probabilities over the set of possible answers,

$$\mathbf{p}_{QI} = W_{QI}\left(\tanh(W_Q\mathbf{x}_Q) \odot \tanh(W_I\mathbf{x}_I)\right). \quad (1)$$

Here, $\mathbf{p}_{QI}$ is the unnormalized probability distribution over the set of all possible answers, given the image, question pair i.e. the model treats the VQA task as a multimodal signal classification task. The answer with the maximum probability is then taken as the predicted answer.

Architecture 2. The second architecture, shown in Figure 2 (right), proposed by Ren et al. [26], borrows ideas from image captioning literature. It treats the image as the first word of the question, by projecting the image feature vector $\mathbf{x}_I$ to the word embedding space with a learnt projection matrix W_e. Following the image first, the question words are then passed one at a time to the LSTM. The hidden state vector of the LSTM after the last time step, which now becomes the joint embedding of the question and the image, is then projected to the answer space to obtain the probabilities over the set of answers, similar to Architecture 1 above.

3.2. Inducing novel objects using auxiliary datasets

Given the above two architectures, we now explain how we introduce novel objects using auxiliary datasets. We experiment with two different settings, first, when the novel words are known textually, and, second, when the novel words are not known. The former is similar to the zero-shot classification [14] setting where the unknown classes are never seen visually at training but are known textually. In the latter, we make the assumption that the novel words are semantically close to the known words; where we use the vector similarity of the words in a standard distributed word embedding space, e.g. `word2vec` [22]. Given the novel words from the two settings, to make the system aware of novel concepts, we have two sources of auxiliary information. We could use large amount of text data, e.g. from Wikipedia or books, as well as image data from large datasets such as ImageNet [28]. We now describe the different ways in which we propose to exploit such auxiliary datasets for making the above described VQA systems aware of novel objects.

Auxiliary text data only. In the first method we propose to use only auxiliary text data for improving VQA performance for novel objects. In most of the VQA architectures, the question encoding is done with a recurrent neural network such as the LSTM network. When large amount of text data is available, which contains both the known and novel objects and the relations between them (as could be described textually), we hypothesise that pre-training question encoder on the auxiliary dataset could be beneficial. To pre-train the question encoder, we use an LSTM based sequence autoencoder (AE), e.g. [6, 16]. The AE is pre-trained on a large external text dataset, e.g. BookCorpus [42]. Figure 3 illustrates the AEs (with the dashed block absent, we explain it more below).

However, this is not a straightforward pre-training as the text vocabulary needs to be expanded to contain the novel words, so that the VQA system is aware of them and does not just see them as `UNK` (special token for all words not in vocabulary). It could be argued that pre-training with only the current vocabulary may improve the encoder in general and might help the VQA system—we test this system as well in the experiments. To do such vocabulary expansion is non-trivial; we could use a vocabulary from the external corpus[1], but such a vocabulary turns out to be very large and can degrade the VQA performance. Thus, we evaluate two ways to construct the vocabulary.

Oracle setting. First, we assume an oracle setting where we know (textually) the novel words that will appear—this is similar to the assumption in zero-shot setting[2] [14]. We add the known novel words to the current vocabulary and train the AE on the auxiliary text data. Once trained, we take the encoder weights from the AE to initialize the question encoder in the VQA system.

General setting. Second, we assume that the novel words would be semantically similar to the known words and, thus, expand the vocabulary by adding words, from the external dataset, which are within a certain distance to the known words. The semantic word distance we use is the cosine distance between the `word2vec` embeddings [22] of the known and novel words. This is a more relaxed as-

[1] All words with frequency above a threshold in the whole dataset

[2] Note that our setting is harder than zero-shot setting in [14] as here the test set contains both the known and novel objects

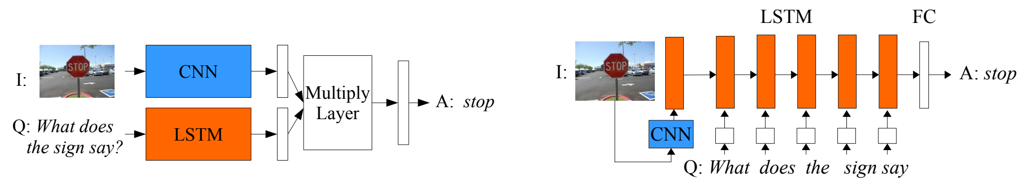


Figure 2: The two Visual Question Answering (VQA) architectures used.

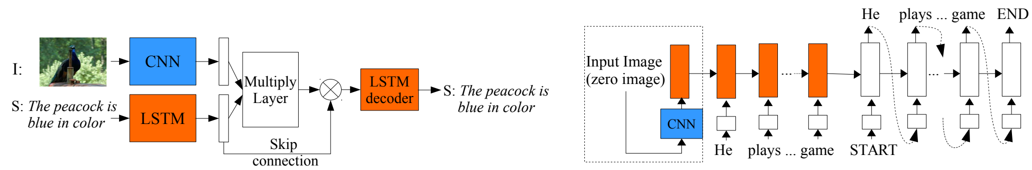


Figure 3: Autoencoders used to pre-train the respective VQA networks' encoders.

sumption compared to the oracle setting and we call this the general setting.

In practice, however, we found that the direct AE training was noisy in this general setting as the vocabulary size increased by nearly $4\times$. We found that the noise and instability of the the training mainly came from the word embeddings, i.e. the projection of the one-hot word representations before being fed to the recurrent unit, in the AE. In order to train it more effectively, we thus use a pre-training technique for initializing the word embeddings of the AE as follows. We first train the AE on the BookCorpus with the VQA vocabulary. We then take the words which appear both in the VQA vocabulary and the original trained `word2vec` vocabulary. Using these words, we obtain a projection to align the the `word2vec` vector space with the currently learned word embedding space. Formally, denote the `word2vec` embedding matrix as A_w and the VQA word embedding matrix as A_v, then we find a projection matrix M, using least squares, as

$$A_w M = A_v, \text{ i.e., } M = (A_w^\top A_w)^{-1} A_w^\top A_v. \quad (2)$$

Once the alignment matrix M is available, the words in the general setting's vocabulary, which are not in the VQA vocabulary, are computed as

$$\hat{A}_v(w) = A_v(w) M, \ \ w \in \mathcal{V}_g \backslash \mathcal{V}_v, \quad (3)$$

where $\mathcal{V}_g$ is the vocabulary set of the general setting and $\mathcal{V}_v$ is that of VQA vocabulary. We term this as vocabulary expansion from the first vocabulary to the second, similar in spirit to the work by Kiros et al. [12]. Going a step further, we initialize the word embedding matrix in the AE using the estimated word vectors and train the AE again on BookCorpus, which we finally use with the base architecture as in the other AE's above.

Auxiliary text as well as image data. Apart from using text only data to induce novel objects, we also attempt to use freely available auxiliary image classification data, e.g. ImageNet [28]. The general philosophy stays the same, we wish to train auto-encoders with the auxiliary data, but in this case such AE takes multimodal input in the form of both text sentences and images, and decodes them back to the sentences. We hope that such an AE[3] will help induce novel objects. To do so, we require paired image-text data and we use the two auxiliary datasets to generate such paired data synthetically and weakly as follows. We take images of the words corresponding to objects in our text vocabulary from the classification dataset such as ImageNet [28] and pair them with general sentences about the object from the text dataset, e.g. BookCorpus [42] or Wikipedia. Note that this is expected to be a noisy paired data; we evaluate if such noise is tolerated by the AE to still give some improvement on the VQA task by learning lexico-visual associations for novel objects.

Since the question encoder for the first architecture does not use the image as an input, we design the corresponding multimodal AE as shown in Figure 3 (left). We take the output of the multiply layer and use it to initialize the hidden state of the decoder. To keep the architecture consistent with the text AE, we introduce a skip connection which feeds the final hidden state of the encoder to the decoder's initial state. Adding such a skip layer ensures that the AE will use the image encoding only if it is beneficial and we hope that this will add resilience to the noise in the synthetically generated paired data. We, thus, effectively sum the final hidden state of the encoder and the output of the multiply layer to obtain the initial decoder state. In case of the second architecture, we just use the image encoding as the first input to the LSTM based AE, as shown in the dashed part of Figure 3 (right).

4. Experiments

We now describe the experiments we performed to validate the method and study VQA when novel objects are present in the test set. We first describe the datasets we used, followed by the new split we created to have novel objects in the test set. We then give our quantitative and qualitative results, with discussions.

VQA dataset [3] is a publicly available benchmark which consists of images obtained from the MSCOCO dataset [17]

[3]It is not strictly an AE as it is only decoding back the text part and not the image part. We refrained from decoding back the images, as initial results were not encouraging; also, image generation from encoded vectors is a complete challenging problem in itself [10].

VQA dataset	
#images	204,721
#ques	614,163
#ans per ques	10
#ques Types	more than 20
#words per ans	one or more

BookCorpus	
#books	11,038
#sentences	74,004,228
#unique words	984,846,357
avg #words / sent.	13

Table 1: Statistics of datasets used

and an abstract scenes dataset. The statistics of the dataset are shown in Table 1. The models are evaluated on the VQA dataset using the accuracy metric defined as

$$\text{acc} = \min\left(\frac{\text{\# humans that provided that answer}}{3}, 1\right). \quad (4)$$

BookCorpus [42] dataset has text extracted from $11,038$ books available on the web. Summary statistics of the dataset are shown in Table 1. We created a split consisting of $73,874,228$ training, $30,000$ validation and $100,000$ test sentences to train the AEs.

ImageNet dataset from the ILSVRC challenge [28] consists of images collected from Flickr and other search engines. Each image is labelled with the presence or absence of one out of 1000 object categories. The training set consists of 1.2 million training images, $50,000$ validation images and $100,000$ test images. We have used ImageNet to obtain images for the known and unknown objects.

Wikipedia. The text data obtained from BookCorpus did not have sentences containing some of the novel objects. Also, the data obtained from BookCorpus was story oriented and not factual data, hence the sentences containing certain objects did not describe the objects themselves, but just contained the objects as a part of a narrative. In order to complement the data from BookCorpus and obtain descriptive information about novel objects, we queried Wikipedia[4] by searching for sentences containing the novel objects.

Weak paired training data. To generate synthetic paired data, we consider all the objects from the oracle/general vocabulary and find an intersection with the ImageNet classes. For each of the objects, we obtain m random images from the matched classes and n random sentences containing the object from BookCorpus and pair them to obtain mn sets of paired images and sentences. In our case, we selected $m = 20$ and $n = 20$. This constitutes the weak paired training data which amounted to approximately 0.25 million samples for the oracle case and 0.45 million samples for the general case.

4.1. Proposed Novel Split for VQA dataset

We create a new split of the VQA dataset to study the setting of novel objects at test time. We obtain the train and validation split of the real scenes part of VQA dataset [3] and call this the original split. The questions from the train split are used for training and the questions from the validation split are used for testing. Next, we divide the full set of images, train and validation combined, into new train and test split as follows. For each of the questions in the VQA dataset, we identify the nouns[5] and create a histogram of the types of questions each noun occurs in. We use normalized histograms to cluster the nouns into 14 clusters. We select 80% of the nouns as known and 20% of the nouns as novel, randomly from each of the 14 clusters. A question in the VQA dataset belongs to the new test set if and only if at least one of the novel nouns occur in it. We randomly sample 5000 questions from the train split to create the validation split. The statistics of the original split and the new proposed are shown in Table 2—note that, while the original test split contains 3178 known objects out of 3330 total, the proposed test split has only 2216 known objects out of a total of 3027, i.e. 811 objects that appear in the test split were never seen (visually or textually) in the VQA training data [6]. Further, Table 3 shows the number of questions where 0 to 5 known objects appear as well (in addition to at least one unknown object). We see that a large number of question, i.e. 32452 contain only novel objects.

	# Questions			# Objects		
Split	Train	Val	Test	Train	Test	Both
Orig	215375	0	121509	3625	3330	3178
Prop	224704	5000	116323	2951	3027	2216

Table 2: Statistics of the dataset splits. The proportion of seen test objects is 95.4% in original vs. 73.2% in proposed.

#Known objs	0	1	2	3	4	5
#Questions	32452	35300	12593	2605	501	48

Table 3: The number of questions with specific number of known words in test set.

Architecture 1	Open Ended Questions				Multiple Choice Questions			
Split	Ov.all	Oth.	Num.	Y/N	Ov.all	Oth.	Num.	Y/N
Orig	54.23	40.34	33.27	79.82	59.30	50.16	34.41	79.86
Novel	39.38	23.07	27.52	74.02	46.54	34.91	29.39	74.10
Drop	**14.85**	**17.27**	**5.75**	**5.8**	**12.76**	**15.25**	**5.02**	**5.76**

Architecture 2	Open Ended Questions				Multiple Choice Questions			
Split	Ov.all	Oth.	Num.	Y/N	Ov.all	Oth.	Num.	Y/N
Orig	48.75	33.31	31.42	74.20	54.94	45.24	32.95	75.28
Novel	34.97	16.98	28.27	71.06	42.83	30.16	29.42	71.12
Drop	**13.78**	**16.33**	**3.15**	**3.14**	**12.11**	**15.08**	**3.53**	**4.16**

Table 4: The drop in performance for novel word setting.

Implementation details. In the case of Architecture 1, we used the default settings of 200 dimensional word encoding size, 512 RNN hidden layer size and 2 RNN layers for computing the results on the case of training only with VQA dataset. To avoid very long training times, with the

[4]Source: https://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2

[5]We used NLTK's PerceptronTagger for obtaining the nouns http://www.nltk.org/_modules/nltk/tag/perceptron.html

[6]The design of the dataset leads to sharing of images between the train and test splits; see supplementary material for detailed discussion.

above large parameter values for other architectures, we selected 512 dimensional word encoding, 512 RNN hidden layer size and 1 RNN layer for computing all our results. We observed that this did not affect our results appreciably. Similarly, in Architecture 2, we used 512 dimensional word encoding, 512 RNN hidden layer size and 1 RNN layer throughout all our experiments.

4.2. Quantitative Results

Our overall results for the two architectures are shown in Table 5. The results are split into the standard question types of Overall, Others, Numbers and Yes/No. We also introduce the Novel question type which consists of all the questions which contain only the novel objects and no known objects (32452 questions from Table 3). This helps us analyse the performance of novel object VQA without interference from the known objects. The image feature, auxilliary data and the vocabulary used for each of the experiments has been specified. The image feature can be `VGG`, `INC` (Inception), `EF` (Early fusion of VGG, INC) or `LF` (Late fusion of VGG, INC), the auxilliary data can be `none` (baseline), `text` (BookCorpus pre-trained AE) or `text+im` (BookCorpus + WeakPaired data pre-trained AE) and the vocabulary can be `train` (only words from train data of novel split), `oracle` (oracle case), `gen` (general case) or `gen(exp)` (vocabulary expansion in general case). We analyse our results in terms of the need to incorporate novel words, effects of different features, vocabulary expansion and pre-training methods on the overall performance. In the following, we refer to a cell in the tables with the Architecture number, the row number and the type of questions (others, numbers etc. in Open Ended or Multiple Choice questions). If we do not specify the sub-type of questions for OEQ or MCQ, then we are discussing the overall averages for these two types.

Performance on original vs. novel split. Table 4 gives the results of the two architectures on the original and novel splits, respectively, without using any data outside of the VQA dataset[7]. We obseve a severe drop in perfomance, e.g. Architecture 1 (2) drops by 27% (28%) on average for the open ended questions, and 21% (22%) on the multiple choice ones. This highlights the fact that the current methods are not capable of generalizing on VQA to novel objects when not explicitly trained to do so. This empirically verifies the argument that VQA in the novel object setting is a challenging problem and deserves attention on its own.

Naïve pre-training is not sufficient. An obvious first argument, as discussed in Sec. 3.2, is that pre-training the text model on the large amount of auxiliary text data, might make it better and hence lead to improved performance, even when the vocabulary is kept the same as the original one (which does not contain the novel words). We tested this hypothesis and found it to not be true. While text only pre-training (rows A1.b, A2.b in Table 5) provided some improvements over the baseline (row A1.a) in most cases, e.g. 39.38 to 40.09 in A1.b OEQ, 46.47 to 47.01 in A1.f MCQ, they were generally minor, especially in the high performing models; some isolated larger improvements did happen, e.g. +6.6% (34.97 to 37.30) in A2.b OEQ, but they were not consistent and happened in the relatively low performing cases only. However, the text only pre-trained models *with* the oracle and general vocabularies provided consistent improvements, e.g. +2.7% (39.38 to 40.44) in A1.c OEQ, +2.3% (40.27 to 41.19) A1.g OEQ, +7.7% (34.97 to 37.68) in A2.c OEQ and +2.3% (37.66 to 38.53) in A2.g OEQ, as they were capable of understanding novel objects. Hence, we conclude that simple pre-training without adding the novel objects to the vocabulary is not sufficient for novel object test setting in VQA.

In the following, all the discussion are w.r.t. methods using vocabularies incorporating novel objects.

Comparison of architectures. We found that that Architecture 1 generally performed better than Architecture 2, e.g. 39.38 on A1.a OEQ vs. 34.97 on A2.a OEQ, 39.56 in A1.k OEQ vs. 35.65 on A2.k OEQ. The relative improvements obtained with the better performing architecture over the corresponding baseline were, unfortunately, generally lesser, e.g. +6.2% and +4.2% in A1.i OEQ and MCQ vs. +11.1% and +8.5% in A2.i OEQ and MCQ, both with early fusion of VGG and Inception features, respectively, indicating that it is more difficult to improve performance for more saturated methods. We do, however, see consistent improvements in majority of cases for both the Architectures, supporting the proposed method.

Auxiliary text data. The models initialized from auxiliary text data, with both oracle and general vocabularies, provide significant improvement in the Yes/No, e.g. +5.6% (71.06 to 75.06) in A2.c OEQ, +4.4% (73.25 to 76.49) in A1.q OEQ, and Novel questions, such as 2.5% (48.03 to 49.23) in A1.g OEQ, +5.2% (44.60 to 46.93) in A2.c OEQ. The proposed model improves on Yes/No questions as they generally have a central object, e.g. 'is the little dog wearing a necktie?' (Fig. 4, image on left-top), and when this object (necktie here) is unknown the baseline model fails to understand the question. Similar trend is visible in the 'Novel' type.

The effect of the general (automatic) vocabulary expansion technique is similar to the oracle case, where the novel objects are assumed to be known *a priori*. The overall results with oracle vocabulary vs. general vocabulary are sim-

[7] While the training/testing data are not same, and hence the performances are not directly comparable, we note that the amount of training data is $\sim 4\%$ more for the models trained in the novel setting (Table 2). If the difficulties of the settings were similar, the novel models should have, arguably, done better due to access to more training data.

ilar, i.e. 41.84 vs. 41.82, 48.87 vs. 48.35, for A1.(i,s) OEQ and MCQ, respectively, and 39.49 vs. 39.91 and 46.40 vs. 46.99 for A2.(i,t) OEQ and MCQ, respectively. Thus, we conclude that the proposed method is capable of leveraging auxiliary text data to improve novel object VQA, in the automatic setting when the minimal assumption is made that the novel words are expected to be semantically similar to the known words.

Vocabulary expansion. Generally, the accuracy of the system improves with vocabulary expansion on the Yes/No and Novel question categories when compared to the accuracy of the non-expanded setting, e.g. 75.48, 48.78 in A1.p OEQ vs. 76.49, 49.36 in A1.q OEQ and 74.38, 51.29 in A2.l MCQ vs. 75.28, 52.47 in A2.m MCQ. This follows the trend from auxiliary text data where we observed similar improvements, and is expected since vocabulary expansion is simply a better way to perform text only pre-training.

Auxiliary text and image data. Using both auxiliary datasets of text and image, as proposed, led to consistent but small improvements over using only auxiliary text datasets. As an example, consider Inception features for Architecture 2 in A2.f–h OEQ. The baseline of 37.66 is improved to 38.53 (+2.3%) by oracle vocabulary expansion and use of auxiliary text data which is further improved to 38.75 (+2.9%) when using both auxiliary data of both text and image—the major improvement comes from using text data and a further small improvement is achieved by using image data as well. We believe that since the text data is relatively clean and rich, it provides good semantic ground for the model to understand the novel objects, while the noisy method of generating weak text-image paired data as proposed is not able to supplement it significantly, and sometimes even deteriorates it slightly. Also, since the image model may have seen the novel objects *a priori*, this may not have a significant impact on the overall results.

Additional observations. Apart from the above main observations, we found that Inception features were generally better than VGG features for VQA. However, most of the improvement of Inception over VGG features was in the "others" category, e.g. 23.07 in A1.a OEQ vs. 24.54 A1.e OEQ and 30.74 in A2.k MCQ vs. 31.87 in A2.o MCQ. The Inception baseline models do not generally perform better than VGG baseline models on the "Novel" questions, especially in Architecture 1 which is the stronger architecture. Therefore, improving image features alone is not sufficient for better novel objects based VQA. This is expected since the text model is still the same and without improvements in the text model or better joint modelling, we cannot expect a significant difference in performance on novel objects.

4.3. Qualitative results

Figure 4 shows some example images with the questions and their answers from the different methods. In the first row, we can observe that proposed model (corresponding to A1.s) has successfully induced the concepts of mouse, apple, event and flavor into the VQA framework whereas the baseline (corresponding to A1.k) has failed to reason based on them. Some of the failure cases of the proposed model are illustrated in the second row. It has failed to induce the concepts of restaurant and direction. We also feel that in the last 2 cases, it has predicted purely based on the text modelling. For example, in the fourth case it says that the plane is in motion and in the fifth case it says the ink is red. This could be because it witnessed similar textual examples and the image is not convincing enough for it to say otherwise.

5. Conclusion

We presented a new task of VQA based on novel objects which were not seen during training. This is a relevant setting as in real world, owing to the heavy tailed distribution of the visual categories, many rare objects are not expected to have annotations. We showed that this is a challenging scenario and directly testing the models which had not seen the objects during training leads to substantial degradation in performances of up to 28%. We proposed to use auxiliary datasets of text, e.g. books and Wikipedia, and images, e.g. ImageNet, to make the system aware of the novel objects it might encounter during testing. We showed that increasing the vocabulary, to include possible novel words, is important and a simple pre-training on the auxiliary data is not sufficient. We proposed two methods for incorporating novel objects in VQA systems. In the first oracle method, we assumed that the novel objects that would appear are given to us, while in the second we made the weaker assumption that the novel words will be semantically similar to the known words. We also proposed a method to use external labeled image datasets to form noisy image-text pairs for pre-training the VQA architectures. Our results demonstrated that making the model aware of novel objects using vocabulary expansion and pre-training on external text datasets significantly improves the performance for VQA in novel test object setting e.g. by +3.4% on the Yes/No questions, +3.6% on Numbers, +11.48% on Others and +4.8% on Novel for Architecture 1 and by +6.76% on the Yes/No questions, +2.2% on Numbers, +24.4% on Others and +8.7% on Novel for Architecture 2 in OpenEnded questions category. However, the gains from external image datasets were either absent or were only modest. We believe that the external text datasets provided a clean and rich source of knowledge while the paired image information was noisy and hence relatively less effective.

Acknowledgements

GS acknowledges support via the Early Career Research Award from SERB India (File # ECR/2016/001158) and the Research-I foundation at IIT Kanpur.

Figure 4: Qualitative examples highlighting the success and failure cases of our proposed model(P) cf. the baseline model(B) and the ground truth(GT). The novel concepts are underlined in the question.

Architecture 1 (A1)

				Open Ended Questions (OEQ)					Multiple Choice Questions (MCQ)				
Row	Feat	Aux	Vocab	Overall	Others	Numbers	Yes/No	Novel	Overall	Others	Numbers	Yes/No	Novel
a	VGG	none	oracle	39.38	23.07	27.52	74.02	47.56	46.54	34.91	29.39	74.10	52.32
b	VGG	text	train	40.09	23.46	28.85	75.14	48.75	47.22	35.32	20.36	75.21	53.39
c	VGG	text	oracle	40.44	23.42	28.24	76.52	48.95	47.65	35.39	29.89	76.60	53.77
d	VGG	text+im	oracle	40.49	23.35	28.32	76.79	48.89	47.38	34.76	30.04	76.87	53.80
e	INC	none	oracle	40.27	24.54	28.02	73.95	48.03	46.47	34.84	29.41	74.00	52.19
f	INC	text	train	40.18	24.12	28.25	74.37	48.10	47.01	35.43	29.91	74.46	52.80
g	INC	text	oracle	41.19	24.98	28.44	75.93	49.23	47.87	36.00	30.24	76.04	53.88
h	INC	text+im	oracle	40.73	24.12	27.80	76.03	48.61	47.23	34.99	29.58	76.12	53.18
i	EF	text	oracle	41.84	25.69	27.93	76.87	49.76	48.47	36.62	29.75	76.96	54.40
j	LF	text	oracle	41.46	25.39	28.66	75.95	49.32	48.22	36.33	30.26	76.54	54.04
k	VGG	none	gen	39.56	23.18	28.47	74.06	48.02	46.23	34.27	29.92	74.13	52.44
l	VGG	text	gen	40.53	23.62	28.93	76.20	49.00	47.50	35.26	30.10	76.27	53.45
m	VGG	text	gen(exp)	40.76	23.89	28.19	76.69	49.05	47.82	35.67	29.40	76.79	53.75
n	VGG	text+im	gen(exp)	40.34	23.09	29.25	76.49	49.25	47.36	34.82	30.31	76.60	53.92
o	INC	none	gen	40.25	24.86	28.12	73.25	47.77	46.53	35.28	29.56	73.33	52.07
p	INC	text	gen	40.76	24.54	28.14	75.48	48.78	46.87	34.77	28.95	75.56	52.83
q	INC	text	gen(exp)	41.39	24.96	28.83	76.49	49.36	47.88	35.74	30.09	76.62	53.77
r	INC	text+im	gen(exp)	40.42	23.77	27.98	75.88	48.77	46.87	34.53	29.11	75.99	52.87
s	EF	text	gen(exp)	41.82	25.72	28.51	76.55	49.60	48.35	36.57	29.81	76.65	53.92
t	LF	text	gen(exp)	39.66	24.03	27.65	73.07	47.34	47.26	35.37	29.42	75.53	53.13

Architecture 2 (A2)

				Open Ended Questions (OEQ)					Multiple Choice Questions (MCQ)				
Row	Feat	Aux	Vocab	Overall	Others	Numbers	Yes/No	Novel	Overall	Others	Numbers	Yes/No	Novel
a	VGG	none	oracle	34.97	16.98	28.27	71.06	44.60	42.83	30.16	29.42	71.12	49.38
b	VGG	text	train	37.30	19.50	26.24	74.48	45.71	44.30	31.26	27.09	74.55	50.31
c	VGG	text	oracle	37.68	19.50	28.28	75.06	46.93	45.12	31.91	29.64	75.11	51.67
d	VGG	text+im	oracle	38.06	20.15	28.45	74.98	47.54	45.80	32.96	30.30	75.10	52.66
e	INC	none	oracle	37.66	20.18	28.32	73.69	46.50	44.59	31.77	29.32	73.77	50.98
f	INC	text	train	37.37	20.00	25.90	73.89	45.54	44.40	31.83	26.59	73.96	50.27
g	INC	text	oracle	38.53	20.79	28.07	75.39	47.55	45.85	32.98	29.37	75.49	52.32
h	INC	text+im	oracle	38.75	21.12	28.96	75.20	47.95	46.07	33.32	30.13	75.34	52.53
i	EF	text	oracle	38.85	21.18	28.43	75.57	48.00	46.47	33.76	30.58	75.66	53.15
j	LF	text	oracle	39.49	22.02	28.71	75.95	48.47	46.40	33.56	29.56	76.04	52.86
k	VGG	none	gen	35.65	17.33	26.62	73.14	45.19	43.64	30.74	27.40	73.28	50.29
l	VGG	text	gen	37.66	19.95	27.73	74.31	46.64	44.99	32.19	29.01	74.38	51.29
m	VGG	text	gen(exp)	38.00	20.21	26.77	75.21	46.84	45.96	33.32	29.26	75.28	52.47
n	VGG	text+im	gen(exp)	37.92	20.21	27.90	74.59	45.58	45.58	33.04	28.99	74.67	52.15
o	INC	none	gen	37.29	19.59	28.76	73.50	46.16	44.63	31.87	27.40	73.28	50.29
p	INC	text	gen	38.23	20.89	28.11	74.22	46.94	45.23	32.54	29.55	74.31	51.67
q	INC	text	gen(exp)	37.99	20.59	26.30	74.65	46.31	45.89	33.54	29.01	74.72	51.84
r	INC	text+im	gen(exp)	38.20	20.49	27.79	75.00	46.97	45.65	32.94	28.96	75.08	51.93
s	EF	text	gen(exp)	38.37	21.13	28.82	74.00	47.10	45.46	33.00	29.99	74.05	51.89
t	LF	text	gen(exp)	39.91	22.75	28.90	75.87	48.48	46.99	34.55	30.24	75.94	53.26

Table 5: Perfomances of the different models in the novel object setting for VQA

References

[1] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Deep compositional question answering with neural module networks. *CoRR*, abs/1511.02799, 2015. 2

[2] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Learning to compose neural networks for question answering. *CoRR*, abs/1601.01705, 2016. 2

[3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. 1, 2, 4, 5

[4] A. Bordes, N. Usunier, S. Chopra, and J. Weston. Large-scale simple question answering with memory networks. *CoRR*, abs/1506.02075, 2015. 2

[5] K. Chen, J. Wang, L. Chen, H. Gao, W. Xu, and R. Nevatia. ABC-CNN: an attention based convolutional neural network for visual question answering. *CoRR*, abs/1511.05960, 2015. 2

[6] A. M. Dai and Q. V. Le. Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems*, pages 3079–3087, 2015. 2, 3

[7] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, 2013. 2

[8] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *CoRR*, abs/1606.01847, 2016. 2

[9] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are you talking to a machine? dataset and methods for multilingual image question. In *Advances in Neural Information Processing Systems*, pages 2296–2304, 2015. 2

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. 4

[11] S. Huang, M. Elhoseiny, A. M. Elgammal, and D. Yang. Learning hypergraph-regularized attribute predictors. In *CVPR*, 2015. 1, 2

[12] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302, 2015. 4

[13] A. Kumar, O. Irsoy, J. Su, J. Bradbury, R. English, B. Pierce, P. Ondruska, I. Gulrajani, and R. Socher. Ask me anything: Dynamic memory networks for natural language processing. *CoRR*, abs/1506.07285, 2015. 2

[14] C. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. In *TPAMI*, 2013. 1, 2, 3

[15] H. Larochelle, D. Erhan, and Y. Bengio. Zero-data learning of new tasks. In *AAAI*, 2008. 1, 2

[16] J. Li, M.-T. Luong, and D. Jurafsky. A hierarchical neural autoencoder for paragraphs and documents. *arXiv preprint arXiv:1506.01057*, 2015. 2, 3

[17] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 4

[18] J. Lu, X. Lin, D. Batra, and D. Parikh. Deeper lstm and normalized cnn visual question answering model. https://github.com/VT-vision-lab/VQA_LSTM_CNN, 2015. 2, 3

[19] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. *CoRR*, abs/1606.00061, 2016. 2

[20] M. Malinowski and M. Fritz. Towards a visual turing challenge. In *NIPS Workshop on Learning Semantics*, 2014. 2

[21] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. oral, to appear. 1, 2

[22] T. Mikolov and J. Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 2013. 3

[23] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013. 1, 2

[24] H. Noh, P. Hongsuck Seo, and B. Han. Image question answering using convolutional neural network with dynamic parameter prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[25] J. Pennington, R. Socher, and C. D. Manning. GloVe: Global vectors for word representation. In *EMNLP*, 2014. 1

[26] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. In *Advances in Neural Information Processing Systems*, pages 2953–2961, 2015. 2, 3

[27] M. Ren, R. Kiros, and R. Zemel. Image question answering: A visual semantic embedding model and a new dataset. *CoRR, abs/1505.02074*, 7, 2015. 2

[28] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 3, 4, 5

[29] K. J. Shih, S. Singh, and D. Hoiem. Where to look: Focus regions for visual question answering. In *Computer Vision and Pattern Recognition*, 2016. 2

[30] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus. Weakly supervised memory networks. *CoRR*, abs/1503.08895, 2015. 2

[31] P. Wang, Q. Wu, C. Shen, A. van den Hengel, and A. R. Dick. Explicit knowledge-based reasoning for visual question answering. *CoRR*, abs/1511.02570, 2015. 2

[32] P. Wang, Q. Wu, C. Shen, A. van den Hengel, and A. R. Dick. FVQA: fact-based visual question answering. *CoRR*, abs/1606.05433, 2016. 2

[33] J. Weston, S. Chopra, and A. Bordes. Memory networks. *CoRR*, abs/1410.3916, 2014. 2

[34] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. v. d. Hengel. Visual question answering: A survey of methods and datasets. *arXiv:1607.05910*, 2016. 2

[35] Q. Wu, P. Wang, C. Shen, A. van den Hengel, and A. R. Dick. Ask me anything: Free-form visual question answering based on knowledge from external sources. *CoRR*, abs/1511.06973, 2015. 2

[36] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele. Latent embeddings for zero-shot classification. In *CVPR*, 2016. 1, 2

[37] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. *CoRR*, abs/1603.01417, 2016. 2

[38] H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. *CoRR*, abs/1511.05234, 2015. 2

[39] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. *arXiv preprint arXiv:1511.02274*, 2015. 2

[40] X. Yu and Y. Aloimonos. Attribute-based transfer learning for object categorization with zero or one training example. In *ECCV*, 2010. 1, 2

[41] Y. Zhu, O. Groth, M. S. Bernstein, and L. Fei-Fei. Visual7w: Grounded question answering in images. *CoRR*, abs/1511.03416, 2015. 2

[42] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *arXiv preprint arXiv:1506.06724*, 2015. 3, 4, 5