Data Augmentation for Visual Question Answering

Kushal Kafle, Mohammed Yousefhussien, and **Christopher Kanan***
Rochester Institute of Technology
{kk6055, mxy7332, kanan}@rit.edu

Abstract

Data augmentation is widely used to train deep neural networks for image classification tasks. Simply flipping images can help learning by increasing the number of training images by a factor of two. However, data augmentation in natural language processing is much less studied. Here, we describe two methods for data augmentation for Visual Question Answering (VQA). The first uses existing semantic annotations to generate new questions. The second method is a generative approach using recurrent neural networks. Experiments show the proposed schemes improve performance of baseline and state-of-the-art VQA algorithms.

1 Introduction

In recent years, both computer vision and natural language processing (NLP) have made enormous progress on many problems using deep learning. Visual question answering (VQA) is a problem that fuses computer vision and NLP to build upon these successes. In VQA, an algorithm is given an image and a question about the image, and it predicts the answer to the question (Malinowski and Fritz, 2014; Antol et al., 2015). Although progress has been rapid, there is still a significant gap between the performance of the best VQA systems and humans. For example, on the open-ended 'The VQA Dataset' that uses real images, the best systems in 2016 are at around 65% accuracy (e.g., Fukui et al. (2016)) compared to 83% for humans (Antol et al., 2015). Analysis of VQA algorithm performance as a function of the amount of training data show that existing algorithms would benefit greatly from more training data (Kafle and Kanan, 2017). One way to address this would be to annotate additional questions about images, but this is time-consuming and expensive. Data augmentation is a much cheaper alternative.

Figure 1: We explore two methods for data augmentation for VQA. The Template method uses semantic image annotations. The LSTM method is a generative approach. For this image, the original questions are: 1) 'Where are the people sitting at?' 2) 'How many orange cups are there?' and 3) 'What is the coffee table made of?' The Template augmentation method generates the questions (4 of 13 total): 1) 'Is there a person in the picture?' 2) 'Is there a couch?' 3) 'How many people are there?' and 4) 'What room is this?' The LSTM method generates the questions: 1) 'How many people are there?' 2) 'How many people are in the picture?' and 3) 'Are they playing a video game?'

Data augmentation is generating new training data from existing examples. In this paper, we explore two data augmentation methods for generating new question-answer (QA) pairs for images. The first method uses existing semantic annotations and templates to generate QA pairs, similar to the method in Kafle and Kanan (2017). The second method is a generative approach using a recurrent neural network (RNN). Fig. 1 shows an example image from 'The VQA Dataset' along with the original questions and the questions generated using our methods. Our methods improve the variety and the number of questions for the image. We evaluate how well each augmentation method performs on two VQA datasets. Our results show that augmentation increases performance for both datasets.

*Corresponding author.

1.1 Related Work

For supervised computer vision problems, e.g., image recognition, labels are scarcer than images. This is especially a problem with deep convolutional neural networks (CNNs) that have millions of parameters. Although more human labeled data would be ideal, it is easier to exploit the training dataset to generate new examples. For image classification, common ways to exploit training images to create more labeled examples include mirror reflection, random crops etc. Many of these methods were used in training the seminal AlexNet (Krizhevsky et al., 2012), which increased the training data by more than ten folds and produced relative improvement of over 4% for image classification.

Compared to vision, where augmentation is common, little work has been done on augmenting text for classification problems. A notable exception is Zhang et al. (2015), where a thesaurus was used to replace synonymous words to create more training data for text classification. However, this augmentation produced little improvement and sometimes even hurt performance. The authors' argued that because large quantities of real data are available, models generalize properly without augmentation. Although *training* using augmented text data is rare, generating new questions about images has been studied. The COCO-QA dataset (Ren et al., 2015) for VQA was created by parsing COCO captions with a syntactic parser, and then used this to create QA pairs for four kinds of questions using hand-crafted rules. However, due to inability of the algorithm to cope with complex sentence structures, a significant portion of COCO-QA questions have grammatical errors or are oddly phrased. Visual question generation was also studied in (Mostafazadeh et al., 2016), with an emphasis on generating questions about images that are beyond the literal visual content of the image. They endeavored to avoid simple questions such as counting and color, which were emphasized in COCO-QA. Unlike our work, their objective was not data augmentation and they did not try to answer the generated questions.

1.2 Datasets and Algorithms for VQA

We conduct experiments on two of the most popular VQA datasets: 'The VQA Dataset' (Antol et al., 2015) and COCO-QA (Ren et al., 2015). 'The VQA Dataset' is currently the most popular VQA dataset and it contains both synthetic and real-world images. The real-world images are from the COCO dataset (Lin et al., 2014). All questions were generated by human annotators. We refer to this portion as COCO-VQA, and use it for our experiments. COCO-QA (Ren et al., 2015) also uses images from COCO, with the questions generated by an NLP algorithm that uses COCO's captions. All questions belong to four categories: object, number, color, and location.

Many algorithms have been proposed for VQA. Some notable formulations include attention based methods (Yang et al., 2016; Xiong et al., 2016; Lu et al., 2016; Fukui et al., 2016), Bayesian frameworks (Kafle and Kanan, 2016; Malinowski and Fritz, 2014), and compositional approaches (Andreas et al., 2016a,b). Detailed reviews of existing methods can be found in Kafle and Kanan (2017) and Wu et al. (2016). However, simpler models such as linear classifiers and multilayer perceptrons (MLPs) perform only slightly worse on many VQA datasets. These baseline methods predict the answer using a vector of image features concatenated to a vector of question features (Ren et al., 2015; Zhou et al., 2015; Kafle and Kanan, 2016). We use the MLP model to conduct the bulk of the experiments, but we show that the proposed method is also effective on more sophisticated VQA systems like multimodal compact bilinear pooling (MCB) (Fukui et al., 2016).

2 Methods for Data Augmentation

The impact of using data augmentation to improve VQA has not been studied. We propose two methods for generating QA pairs about images: 1) a template based generation method that uses image annotations and 2) a long short term memory (LSTM) based language model. The number of questions generated using both methods are shown in Table 1.

2.1 Template Augmentation Method

The template data augmentation method uses the semantic segmentation annotations in COCO to generate new QA pairs. COCO contains detailed segmentation annotations with labels for 80 objects typically found in the images. We synthesize four kinds of questions from the COCO annotations: yes/no, counting, object recognition, scene, activity and sport recognition.

Yes/No Questions: First, we make a list of the

Table 1: Number of questions in COCO-VQA compared to the number generated using the LSTM and template methods.

Type	COCO-VQA(Antol et al., 2015)	LSTM	Template	Total Augmentation
Yes/No	140,780 (38.0%)	31,595 (29.2%)	1,023,594 (86.2%)	1,055,189 (81.5%)
Number	45,813 (12.4%)	2,727 (2.52%)	60,547 (5.1%)	63,274 (4.8%)
Other	183,286 (49.6%)	73,617 (68.2%)	102,617 (8.6%)	176,234 (13.6%)
Total	369,879	107,939	1,186,758	**1,294,697**

COCO objects present in an image. If the object has an area greater than 2000 pixels, we can generate an object presence question, e.g., 'Is there a OBJECT in the picture?' with 'yes' as the answer. We use 10 templates to allow some variation in phrasing. For example, 'Is there a person in the image?' and 'Are there any people in the photo?' are variations of the same question. To avoid question imbalance, we ask equal number of 'no' questions about the objects that are absent from the image.

Counting Questions: To make counting questions, we count the number of separate annotations of all the objects of a particular category that have an area greater than 2000 pixels, and ask 12 variations of a counting question template.

Object Recognition Questions: Object recognition questions such as 'What is in the picture?' can be ambiguous because multiple objects may be present. So, we ask questions about COCO 'super-categories' (e.g., 'food,' 'furniture,' 'vehicle,' etc.) to specify the type of object in the question. However, ambiguity may persist if there are multiple objects belonging to same supercategory. For example, 'What vehicles are shown in the photo?' becomes ambiguous if both 'cars' and 'trucks' are present. So, we ensure only a single object of a supercategory is present before asking a recognition question. We use 12 variations of 'What SUPERCATEGORY is in the image?'

Scene and Activity Questions: If a object in an image belongs to the COCO supercategory *indoor* or *outdoor*, we generate questions such as 'Is this indoor or outdoors?' Similarly, we ask about different rooms in the house by identifying the common objects in the room. For example, if there are least two common kitchen appliances in the picture(e.g., toaster, microwave, etc.), then we infer the room is a kitchen and ask 'What room is this?' with 'kitchen' as the answer. We employ similar strategies for 'living room' and 'bathroom.' We used six variations for 'indoor/outdoor' questions and four variations for room classification questions. For sports, we check if any sports equipment is present in the image and generate a question about the type of sport being depicted in the image. We use four variations of questions to ask about each of the six common sports activities.

2.2 LSTM Augmentation Method

One major issues with our template-based augmentation method is that the questions are rigid and may not closely resemble the way questions are typically posed in the VQA dataset. To address this, we train a stacked LSTM that generates questions about images. The network consists of two LSTM layers each with 1000 hidden units followed by two fully connected layers, with 7000 units each, which is the size of our vocabulary constructed by tokening training questions into individual words. The first fully connected layer has a ReLU activation function, while the second layer has the 7000-way softmax. The output question is produced one word at a time until the ¡end-of-question¿ token. The network is trained using the COCO-VQA training data. During the generation process, we start by passing the ¡start-question¿ token concatenated with the image features. To predict the next word, we sample from a multinomial distribution characterized by the prediction probabilities. Sometimes such sampling generates questions unrelated to image content. To compensate for this, we repeat the sampling for every word multiple times and pick the word occurring most frequently. We then generate 30 initial questions per image, and only retain the 3 most frequent questions. Any generated question that already exists in the original dataset is removed.

We use the MLP VQA method described in Sec. 3 to create answers for the generated questions, but it is trained without augmented data. Used alone, this can produce many incorrect answers. To mitigate this problem, we tried to identify the kinds of questions the MLP VQA algo-

COCO-VQA: What instrument does the person who lives here play? **A:** Guitar

COCO-QA: What is in front of a computer looking at the screen as if browsing? **A:** Cat

Figure 2: Examples of questions and predicted answers from COCO-VQA and COCO-QA datasets. The results are from model trained jointly on original and temple based QA pairs.

rithm tends to get correct. To do this, we use k-means to cluster the training question features concatenated to a one-hot vector with the answer for each question type ($k = 25$). We assign each validation QA pair to one of these clusters and compute each cluster's accuracy. QA pairs assigned to clusters that have a validation accuracy of less than 70% are removed from the dataset.

3 Experiments And Results

First, we use the simple MLP baseline model used in Kafle and Kanan (2016) to assess the two data augmentation methods. Kafle and Kanan (2016) showed that MLP worked well across multiple datasets despite its simplicity. The MLP model treats VQA as a classification problem with concatenated image and question features given to the model as features and answers as categories. CNN features from ResNet-152 (He et al., 2016) and the skip-thought vectors (Kiros et al., 2015) are used as image and question features respectively. We evaluate the MLP model on COCO-VQA and COCO-QA datasets. For COCO-QA, we excluded all the augmented QA pairs derived from COCO's validation images during training, as the test portion of COCO-QA contains questions for these images. Table 2 shows the results for the MLP model when trained with and without augmentation. Some examples for the model trained with augmentation are are shown in Fig. 2.

Next, to demonstrate that the data augmentation scheme also helps improve more complex models, we train the state-of-the-art MCB model with attention (MCB+Att.+GloVe) (Fukui et al., 2016) with the template augmentation and compare the accuracy when the same model trained only on the COCO-VQA dataset (Table 3).

Table 2: Results on COCO-VQA (test-dev) and COCO-QA datasets for the MLP model trained with and without augmentation.

Method	COCO-QA	COCO-VQA
MLP (Our Baseline)	60.80	58.65
MLP (LSTM Augmentation)	61.31	58.11
MLP (Template Augmentation)	62.21	**59.61**
MLP (Joint Augmentation)	**62.28**	58.45

Table 3: Results on COCO-VQA (test-dev) for the MCB+Att.+GloVe model trained with and without template augmentation.

Method	COCO-VQA
MCB+Att.+GloVe	64.7
MCB+Att.+GloVe (Template Aug.)	**65.28**

4 Discussion and Conclusion

Referring to Table 2, we can see that both forms of augmentation improved accuracy on COCO-QA compared to the baseline, and the template-based approach worked better than LSTM. For COCO-VQA, the template-based augmentation helped considerably, producing a relative increase of 1.6% compared to when it was not used. We did not observe an improvement from using the LSTM method, perhaps due to label noise. While we tried to mitigate label noise by rejecting QA pairs that were likely to be wrong, this was not sufficient. We are exploring alternative training methods that are robust to label noise (e.g., Reed et al. (2014)) to help improve results using LSTM.

Additionally, we also evaluated which types of questions benefit the most from data augmentation. For the MLP model trained on COCO-VQA with the template augmentation, counting category answer is improved the most (1.74%), followed by others (1.01%), and yes/no (0.7%).

The results are promising and demonstrate that VQA algorithms can benefit from data augmentation, even for hard question types like counting. Furthermore, there is a lot of room for expansion in both the LSTM and the template based methods to produce a larger number and variety of questions. Template augmentation worked best in our experiments, but if we can control for label noise, the LSTM method can be more flexible than the template method, and could be used to generate virtually unlimited amount of training data using images from the Internet.

References

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016a. Deep compositional question answering with neural module networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016b. Learning to compose neural networks for question answering. In *Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *The IEEE International Conference on Computer Vision (ICCV)*.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Conference on Empirical Methods on Natural Language Processing (EMNLP)*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kushal Kafle and Christopher Kanan. 2016. Answer-type prediction for visual question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kushal Kafle and Christopher Kanan. 2017. Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding* .

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems (NIPS)*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems (NIPS)*.

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems (NIPS)*.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Larry Zitnick, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating natural questions about an image. *CoRR* abs/1603.06059.

Scott E. Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. 2014. Training deep neural networks on noisy labels with bootstrapping. *CoRR* abs/1412.6596.

Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring models and data for image question answering. In *Advances in Neural Information Processing Systems (NIPS)*.

Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2016. Visual question answering: A survey of methods and datasets. *arXiv preprint arXiv:1607.05910* .

Caiming Xiong, Stephen Merity, and Richard Socher. 2016. Dynamic memory networks for visual and textual question answering. In *International Conference on Machine Learning (ICML)*.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. 2016. Stacked attention networks for image question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems (NIPS)*.

Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. 2015. Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167* .