Data Augmentation for Visual Question Answering

Deshana Desai
New York University
dkd266@nyu.edu

Anish Shah
New York University
abs699@nyu.edu

Tushar Anchan
New York University
tra290@nyu.edu

Chhavi Yadav
New York University
cy1235@nyu.edu

Abstract

Visual Question Answering (VQA) systems require understanding of vision, language and common sense knowledge. This is a challenging task with limited available data. Data Augmentation is a technique used widely in various Computer Vision tasks to expand the dataset. This makes the models resistant to over-fitting and helps improve the generalization error. However, Data Augmentation in Natural Language Processing is not as straight-forward. We propose four different techniques for Data Augmentation in VQA. We use existing semantic annotations, information extracted from image features and language features to generate new questions. Our fourth method uses a generative model for Visual Question Generation. We experiment our proposed methods on VQA v2 dataset and evaluate the performance with state-of-the-art VQA algorithms.

1 Introduction

VQA involves answering a question about an image. Although humans can easily answer arbitrary questions about an image, developing models for this task is particularly challenging. A VQA network must learn mappings from the joint space containing image features and textual representations (Questions) to labels (Answers). The first major dataset to be released was DAQUAR (Malinowski and Fritz, 2014). While DAQUAR was a pioneering dataset for VQA, it was too small and biased to successfully train and evaluate more complex models. This demonstrates the importance of a large and diverse dataset for VQA.

An ideal VQA dataset needs to be sufficiently large to capture the variability within questions, images, and concepts that occur in real world scenarios (Kafle and Kanan, 2017). The current state-of-the-art VQA system (Fukui et al., 2016) for the Open Ended VQA Dataset is about 65%, as opposed to the human performance which is 83% (Antol et al., 2015). The difference between the two shows a substantial scope for improvement. It is common knowledge that the more data the network is exposed to, the better is the generalization ability of the network and more the resistance to over-fitting. Further, collecting a large dataset for this task requires crowd-sourcing which is time consuming and expensive. The data is therefore not readily available and is hard to collect. In sight of the above points, we hypothesize that using Data Augmentation for VQA will help the model generalize better and is a cheaper way to create larger datasets. The VQA v2 dataset used contains COCO and abstract scene images, annotations, QA pairs. The goal of this paper is to use Data Augmentation to help expand the VQA v2 dataset and thereby improve the performance of VQA systems.

In summary, we make the following contributions:

- We enumerate the different question types that can be encountered in general VQA datasets.
- We implement the template based augmentation methods described in Kafle and Kanan (2017) using COCO Annotations.
- We describe template based algorithms which extract information from the image and generate questions using this information applied as an operator over the templates. Images contain rich information that can be harnessed to generate a wide variety of questions.
- We present methods for language-only augmentation. These methods ensure the model does not overfit to biases in question phrasing.
- We propose and test models for generating

questions as well as question-answer pairs from existing data using an Attention based LSTM model. We exploit information from images and captions to generate questions that are close to those present in the augmented dataset. The model can generate a wider variety of complex image-specific questions without having to hand-craft templates.
- We provide an empirical evaluation showing relative contributions of proposed augmentation methods on VQA v2 dataset using the strong baseline VQA system Show, ask, attend and answer (Kazemi and Elqursh, 2017).
- We release the code [1] and data from the work.

Our methods are chosen so as to ensure diversity in the generated QA pairs.

2 Related Work

Kafle and Kanan (2017) discussed existing datasets and algorithms for VQA. They analyzed existing algorithms with respect to the training data size and showed that algorithms perform better for larger training size.

Kafle et al. (2017) used this result as a basis to test how different data augmentation techniques affect the VQA task performance. They demonstrated two methods for generating new questions. The first method is a template-based method that uses semantic annotations from MSCOCO dataset (Lin et al., 2014). They generate four types of questions: yes/no, counting, object recognition and scene-activity. The second method is a generative one using an LSTM model. Their template based methods helped improve VQA considerably compared to the model producing an increase of 1.6% from baseline performance on the VQA dataset (Antol et al., 2015). They argued that the LSTM model was not able to improve performance due to large amount of label noise and that the methods for rejecting QA pairs(which were likely to be wrong) were not sufficient. Their experiments concluded that VQA algorithms benefit from data augmentation even for hard question types like counting and there is a lot of room for improvements in the methods used.

Our work delves deeper into the variety of questions that can be generated and produces a larger number of questions. We also do not restrict our data augmentation to the limited kind of questions seen in the VQA 1.0 dataset. Their template based methods directly use the information available in the COCO Annotations for the images. We implement the Template based methods proposed by them using COCO Annotations and additionally, propose Template based methods that extract information to generate questions from image features. Further, we add language-only augmentation methods to combat over-fitting in the question phrasing. Finally, our Attention based LSTM model is aimed at reducing noisy QA pair generation.

Visual Question Generation(VQG) is another very recent and open-ended thread of research in the VQA domain. Ren et al. (2015) proposed a rule-based algorithm to convert a given sentence into a corresponding question that has a single word answer. Mostafazadeh et al. (2016) were the first to learn a VQG model. They focus on creating ”natural” and ”engaging” questions. This was also the first paper to draw a parallel between the task of Image Captioning and Image Question Generation. Their best performing model is based on the state-of-the-art multimodal RNN model used for Image Captioning. We draw inspiration from this work to base our VQG Model from current state-of-the-art techniques used for Image Captioning Mun et al. (2017). A different approach used for this task is using VAE with LSTM networks proposed by Jain et al. (2017). The advantage of this model is the ability to generate a large set of varying questions from the given image. In the same spirit, Li et al. (2017) introduced iQAN which can accomplish VQA and its dual task VQG simultaneously. It achieves this by gradually adjusting its focus of attention guided by both a partially generated question and the answer. It showed improvement in accuracy on VQA v2 (Goyal et al., 2017) and CLEVR dataset (Johnson et al., 2017).

VQA requires the model to learn not only from the given text and image, but also the cross-modal mapping between the two spaces. In sight of this, Ben-younes et al. (2017) introduced MUTAN, a multimodal tensor-based Tucker decomposition which aims at learning an alignment between the visual and textual feature representations. They demonstrate how their model generalizes to latest VQA architectures providing state-of-the-art results. As a data augmentation method, they tripled the size of the training set by using additional data from the Visual Genome dataset (Kr-

[1] github.com/deshanadesai/VQA-DataAugmentation

ishna et al., 2017) to train their model. Our work not only checks performance improvement with the existing crowd-sourced questions from the Visual Genome dataset but also extensively looks at different approaches with which the region graphs from the dataset can be harnessed to produce questions.

We use various methods for language-only augmentation for the QA pairs. A full list of various kinds of possible semantic-preserving paraphrases is given in Bhagat and Hovy (2013). According to their analysis, Synonym Substitution and Function Word Variations are the most commonly encountered lexical changes in a paraphrase corpus.

3 Methods for Data Augmentation

Questions that can be asked to query images in a robust VQA system can be broadly categorized into the types show below.

3.1 Template Augmentation using Annotations

We use annotations from COCO dataset (Lin et al., 2014) to generate new questions. We implement Object Presence, Object Recognition & Counting Questions as done in Kafle and Kanan (2017).

Relative and Absolute Position Reasoning - We use annotations to get bounding boxes, sub and super category of objects. The objects are then arranged in left to right and up to down fashion. We remove ambiguity by considering only the top left of the bounding boxes. Next, we generate a host of questions like "What is the rightmost object?","What is below the table?" , "Is the bicycle to the left of the pot?", etc. We can use different attributes such as color, counting, size of object etc. along with this to generate ensemble type questions like "How many chickens are to the left of the red lamp post?".

Visual Genome Dataset- We exploit this dataset to create questions based on objects, attributes, scene and activity. We also use these annotations to create verbose questions that are not otherwise covered such as "Who is on top of a brown, spotted horse in a green field?"

3.2 Template Augmentation using Image Features

The augmentation methods outlined below extract information from the image and use this with textual templates:

Scene and Activity Recognition We use the Image classification model trained on the Places dataset (Zhou et al., 2017) to get answers. We select particular variations of hand-written templates such as "What room is this?" or "Where is this image from?" depending on the answer. Our second approach is using the Visual Genome dataset (Krishna et al., 2017) annotations to create scene and activity related questions.

Color Based (eg. What color is the car?) We extract the segmented region, category and supercategory of the object from the annotations. We use DBSCAN (Ester et al.) to cluster similar RGB values together. Since we are interested in the dominant colors in the object, we sort the clusters by size and pick the largest ones. Next, we map the median of the picked clusters to the human color space. We filter out objects which have multiple dominant colors to reduce ambiguity.

Descriptive and Sentiment based Questions (eg. What kind of waves are in the ocean? or How is the woman feeling?) - We synthesize questions that address cues about the effect, emotion, and sentiment of the visual content in the images. We use a pre-trained SentiBank concept detector (Borth et al., 2013) which has been trained on a Visual Sentiment Ontology dataset consisting of Flickr images and corresponding Adjective-Noun Pairs(ANP). It uses visual features of the image to train a model to identify the most relevant ANP. We synthesize questions related to the noun with adjectives produced by the model as answers.

3.3 Language-Only Template Augmentation

We use NLTK Python toolkit (Bird et al., 2009) and WordNet (Miller, 1995) for below methods:

Synonyms We filter out "stopwords" and pick the nouns and verbs in the question statement for synonym substitution. We find synsets based on the POS tag. There may be multiple synsets representing multiple contexts the word W may appear in. We use a simple algorithm for Word Sense Disambiguation. We create a Word Sense Profile for each of the senses represented by the synsets (using hypernyms, synonyms, hyponyms, similar words). We also create a word sense profile for W using other tokens in the question and related caption. Finally, we pick the synset which has the best matching word sense profile to that of W.

Semantic preserving paraphrase (eg. Where is the man skateboarding? $\rightarrow$ At which location, is the man skateboarding?) We use four steps to generate a paraphrase for a question. First, we extract the Noun and Verb Phrases from the question. Next, we find the subject, object and verb from these respective phrases. Using this, we convert the question into an immediate query of the form $\phi(q) =< where, man, skateboarding >$. The second step is to use handwritten templates and tagging-based rules to transform the queries into a standard form $\sigma(q) =< location, person, activity >$. Now, we have each question statement mapped to an immediate and standard query respectively. These steps performed a down-mapping from the question space to the query space. The third step is - given a new question statement Q, $\phi(Q)$ and $\sigma(Q)$; we find a candidate pool CP of Questions with the matching $\sigma(Q)$ and use the format of the statements in CP to generate a new set of questions for Q. This step involves an up-mapping from the standard query space back to the question space. In the final step, we filter out noise based on handwritten parsing-based grammar rules.

Converse Questions (eg. Is the woman feeling happy? → Is the woman feeling unhappy?) We use antonym substitution for adjectives to generate converse questions. This also requires the negation of the answer. At the moment, this can only (a) be applied to Yes/No type questions or (b) applied to non open-ended questions to generate Yes/No type QA pairs.

Textual Entailments (e.g. Is the man snoring → Is the man sleeping?) We tokenize a given question and find the corresponding POS tags. Next, we use the WordNet and ConceptNet knowledge base to extract entailment relationships between the list of tokens and the words in the vocabulary.

3.4 Visual Question Generation Models

Visual Question Generation is similar to Image Captioning since both try to model the joint distribution of images and text to generate information. In VQG, information corresponds to the generated question. We generate questions using an Attention based LSTM model inspired by Mun et al. (2017). The model helps in generating image-specific diverse questions without resorting to rule-based algorithms.

Our first proposed model is given as input the image, captions and answer to the question that we wish to generate. We use ResNet (He et al., 2016) to obtain feature vectors of the image. We encode the captions and answers using a pre-trained Skip-Thought Vector (STV) model (Kiros et al., 2015). The last hidden state of the Gated Recurrent Unit (GRU) of the STV model, obtained after passing the caption/answer, is considered to be the embedding vector E_c/E_a respectively. These vectors are concatenated and passed to the attention layer. We use an attention mechanism and train the model to steer its attention to a relevant region and generate a question related to that region. The output of this layer is fed to an LSTM which actually generates the question. Training is done using VQA v2 dataset which has both captions and QA pairs. If the question-answer pair already exists in the dataset, the question is ignored. Similarity between questions is evaluated using the BLEU metric.

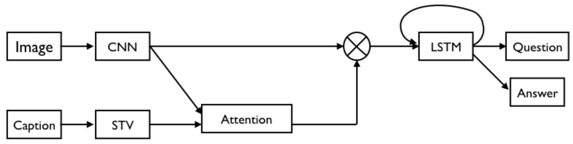

Figure 1: Attention Based LSTM Model

A bottleneck of the above model is that it can only generate a question that corresponds to the answer given as input to it. We propose a variant (1) that generates a Question-Answer pair (q, a). In this case, the inputs to the model are image feature vectors and the caption embeddings. The answer is the last word generated by the LSTM. This can be extended to generate open ended answers. The loss to be minimized is given by (1) and (2)

$$L = -\log p((q, a) \mid f_{att}(I, c)) \tag{1}$$

$$= -\log p(w_1 \mid w_0, f_{att}(I, c)) + \sum_{t=1}^{T} -\log p(w_{t+1}|w_t, h_{t-1}) - \log p(w_{T+2} \mid w_T \dots w_1, h_{T+1}) \tag{2}$$

where q is the question composed of (w_1,w_2,.....w_T), w_0 is the <Beginning of question> tag while w_{T+1} is the <end of question> tag, I is the image, c is the caption, a given by w_{T+2} is the answer, f_{att} is the attention function that calculates the attention and the initial hidden state h_{-1} for the LSTM using I, c only once in the beginning , h_{t-1} is the previous hidden state of the LSTM.

Experiments and Results

Acknowledgement

We would like to thank Sam Bowman, Paloma Jeretic and Nishant Subramani for their inputs. We would also like to thank NYU HPC team for providing us with GPUs.

Collaboration Statement

All four of us brainstormed ideas/resources together and also wrote the partial draft together.

- Deshana Desai - Implementing Language Only Augmentation Methods, Color Based Template Augmentation
- Anish Shah - Reproducing Template-based methods like Object Presence, Object Recognition, Counting from Kafle et al. (2017) and Implementing Visual Genome Dataset related methods
- Tushar Anchan - Implementing Descriptive and Sentiment based questions, Synonyms, Scene and Activity Recognition
- Chhavi Yadav - Implementing relative and absolute position reasoning and Visual Question Generation Models

References

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

Hedi Ben-younes, Rémi Cadène, Matthieu Cord, and Nicolas Thome. 2017. MUTAN: multimodal tucker fusion for visual question answering. *CoRR*, abs/1705.06676.

Rahul Bhagat and Eduard Hovy. 2013. What is a paraphrase? *Computational Linguistics*, 39(3):463–472.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 223–232. ACM.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise.

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 *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 457–468.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Unnat Jain, Ziyu Zhang, and Alexander G. Schwing. 2017. Creativity: Generating diverse questions using variational autoencoders. *CoRR*, abs/1704.03493.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 1988–1997. IEEE.

Kushal Kafle and Christopher Kanan. 2017. Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, 163:3–20.

Kushal Kafle, Mohammed Yousefhussien, and Christopher Kanan. 2017. Data augmentation for visual question answering. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 198–202.

Vahid Kazemi and Ali Elqursh. 2017. Show, ask, attend, and answer: A strong baseline for visual question answering. *arXiv preprint arXiv:1704.03162*.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, and Xiaogang Wang. 2017. Visual question generation as dual task of visual question answering. *arXiv preprint arXiv:1709.07192*.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

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*, pages 1682–1690.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating natural questions about an image. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1802–1813.

Jonghwan Mun, Minsu Cho, and Bohyung Han. 2017. Text-guided attention model for image captioning. In *AAAI*, pages 4233–4239.

Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring models and data for image question answering. In *Advances in neural information processing systems*, pages 2953–2961.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.