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 three 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 experiments on VQA v2 dataset show that augmentation using language features improves performance of the baseline VQA model.

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. Foody et al. (1995) show 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 (Goyal et al., 2017) contains COCO and abstract scene images, annotations, Question-Answer (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 this paper, we explore three template based data augmentation methods for generating new QA pairs for images. Our first method uses information extracted from annotations of the Visual Genome dataset (Krishna et al., 2017). This dataset contains rich information about relative positioning between objects which we harness to generate QA pairs. In addition, images are also a source of rich information that can be harnessed to generate a wide variety of questions. Our second method focuses on extracting color and scene recognition based information from image features. In our final method, we solely focus on augmentation in the textual space. We use augmentation techniques such as Synonym replacement, Converse Questions and Textual Entailment to perturb a QA pair. This can help prevent over-fitting to biases in question phrasing. We evaluate how each augmentation technique performs on VQA v2 dataset using the strong baseline VQA system Show, ask, attend and answer (Kazemi and Elqursh, 2017). Our results show

that Language-Only augmentation techniques improve performance of the baseline model indicating presence of language bias in the model previously. We release the code[1] and data from our work.

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 v1 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.

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. 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 (Krishna et al., 2017) to train their model. Our work 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.

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

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).

Visual Genome The Visual Genome Dataset (Krishna et al., 2017) contains dense image annotations. For each image in the dataset, it has information about the objects, their attributes and relationship between objects in the image. The relationship is represented in the form *<subject, predicate, object>*. For example, *<man, on top of, camel>*, *<camel, on, grass>*. We generate a graph using these relationships. The node is represented by a subject or object and the edge is represented by the predicate. We then select a node and generate questions by depth-first search. For the above relationships, we generate questions such as "Who is on top of the camel on the grass?", "What is on the grass?". We tune the generated questions using Part-of-Speech tagger (Toutanova et al., 2003; Toutanova and Manning, 2000) to make it grammatically correct.

3.2 Template Augmentation using Image Features

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

Scene Recognition (eg. What is the place where this picture is taken? → Restaurant) We use DenseNet161(Huang and Liu, 2017) trained on the Places dataset (Zhou et al., 2017) to generate information regarding the scene in an image. We filter noise by discarding information where the model has a confidence less than 0.4. We use this information as the answer and generate questions based on hand written templates. We use 10 different base templates of the type "What is this *<END>*", "Where is this *<END>*" etc. We replace the *<END>* tag with one of the 30 different templates such as "in this picture" depending on the base template. We attempt to avoid over-fitting to the question phrasing by using multiple possible combinations for the beginning as well as the end of the question.

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.

We generate questions for this method using 2 types of templates. First, We use 5 different base templates of the type: "What is the color of *<OBJ> <END>* ?". We use different variations for substituting *<END>* such as "in this picture". Then, We extend the above method and use 5 different base templates to extract only the dominant color of an object. The templates are of the type: "What color most stands out in the *<OBJ> <END>* ?". If an object has multiple colors, we select the color cluster with the largest number of points.

DBSCAN computes the full $O(n^2)$ distance matrix to cluster points which is computationally expensive. For large objects, this process is highly time consuming. To avoid this, we create a sliding window of 1×1 with a stride of 2 thereby only picking every other pixel in the segmented region. This speeds up computation and does not result in a noticeable difference in the output.

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 rank the synsets in order of best matching word sense profile to that of W. To calculate similarity between the tokens in the word sense profile, we use the Word2Vec model (Mikolov et al., 2013) pre-trained on Google News corpus (3 billion running words) to represent tokens as Vectors and compute the Cosine Distance between these vectors. We pick the best match (synset) among the different possible senses

(synsets). We substitute the given word W with synonyms from this synset. Note that we take into account whether the token to be substituted is in singular or plural form (based on the POS tag) and accordingly modify the substitute word to its plural form if required.

Converse Questions (eg. Is the woman feeling happy? → Is the woman feeling unhappy?) We use antonym substitution for adjectives to generate converse questions. For a given question, we identify the adjectives based on its POS Tag. We choose a corresponding synset based on the Word Sense Disambiguation procedure outlined above. For each lemma in this selected synset, we query the WordNet database for possible antonyms. If antonym is not a single word, it is discarded since it may require a change in the sentence structure. We randomly select one of the possible single word antonyms and use it for substitution. This also requires the negation of the answer. At the moment, this is only applied to Yes/No type questions.

Textual Entailments (eg. Is the man snoring → Is the man sleeping?) We tokenize a given question and find the corresponding POS tags. We check whether the head verb is not negative and only then perform entailment substitutions. For each verb, we check if a corresponding entailment exists in the WordNet database (eg. "snoring" is a piece of text T that entails another text "sleeping"). We also identify if any noun in the sentence has a hypernym suitable for lexical substitution. We employ Word Sense Disambiguation as described above to ensure that the results are less ambiguous.

4 Experiments and Results

Table 1 shows the number of new questions generated through each augmentation technique. We chose to create the following grouping of questions: Object attributes, higher level scene understanding, language perturbations and questions regarding relationships between objects.

In the table, Baseline refers to the number of questions present in the original VQA v2 training dataset. The "Object Attributes" method represents questions generated from the (Kafle et al., 2017) methods and the Color Based method in Section 3.2. The "Language Only" method represents questions from techniques in Section 3.3, "Scene Recognition" methods represents questions from techniques in Section 3.2 and "Visual Genome" method represents questions from techniques in Section 3.1. The "Cumulative" method refers to usage of all the above methods to generate questions.

Table 1: Number of Questions in VQA v2 compared to the number of new Questions generated through different methods to be added to the Questions in the VQA v2 dataset.

Method	# Questions Added
Baseline	443,757
+ Object Attributes	446,660
+ Visual Genome	598,802
+ Language Only	772,759
+ Scene Identification	36,924
+ Cumulative	1,793,697

Table 2: Results on VQA v2 Validation dataset for the Show, Ask, Attend and Answer baseline with and without different Template Based Augmentation Methods.

Method	Accuracy
Baseline (Kazemi and Elqursh, 2017)	59.12
+ Object Attributes	57.89
+ Visual Genome	56.67
+ Language Only	**59.51**
+ Scene Identification	58.82
+ Cumulative	58.67

4.1 Overall Performance

Table 2 shows the performance of our methods against the original baseline performance on the validation set of VQA v2. The Language Only augmentation outperforms other methods as well as the baseline. There is a drop in performance observed for the other methods compared to the baseline. The models were trained for 20 epochs keeping the hyper parameters configurations the same.

4.2 Performance stratified by question types

To gain more insight into the performance of our models, we report its performance across the question types in Table 4.

We notice a significant performance increase in question types - "Do you" (eg. "Do you see a camera?") and "Is there a" which deal with Object presence related questions. Since the Object Attributes methods contain questions regarding Object Presence and Recognition, the improvement

Table 3: Results on VQA v2 Validation dataset stratified by Answer Type.

Method	Yes/No	Num	Other
Baseline	76.90	37.29	51.37
+ Object Attributes	76.46	35.09	50.43
+ Visual Genome	74.97	33.10	48.99
+ Language Only	**77.31**	**37.79**	**51.74**
+ Scene Identification	76.59	**37.41**	50.98
+ Cumulative	**76.99**	35.54	50.96

is reasonable. On the contrary, even though the Object Attributes methods use COCO annotations to produce counting and color type questions, we observe a decrease in performance for the question types "How many", "How many people are in" and "What color".

The accuracy is improved by Language Only augmentation for 78% of the question types. This demonstrates a versatility in boosting performance across different question types. However,we observe a slight decrease in performance for "What number is ", "Do you" and "Which ".

We notice a 0.2% improvement in performance for the question type "What room is " for the Scene Identification method. This aligns with the type of questions created with this method. However, we do not notice any significant effect otherwise of this method.

Visual Genome method produces complex questions regarding relationships between objects. However, the question types "What is " (eg "What is to the left of the bicycle?") and "Where is the" both result in 2% decrease in performance.

We notice that an accuracy increase in all four methods leads to a definite increase in the cumulative accuracy. However, the methods do not seem to complement each other otherwise.

4.3 Performance stratified by answer types

We also report performance of the models across the answer types. This is a shallower classification since we divide the types of answers into "Yes/No", "Number" and "Other". We note that the accuracy increases for all the three answer types only for the Language Only method. It is interesting to note that this method improved performance for Yes/No and Number type answers while Object Attributes did not. We observe a 0.12% increase in the accuracy of Number type answers for Scene Identification and 0.09% increase in the accuracy of Yes/No type answers for Cumulative method.

4.4 Language Only Augmentation

We take a closer look at the Language Only Augmentation methods. The results reported above in Section 4.1 were for **Single Word Substitutions**. We chose tokens not belonging to Stop Words and substituted them based on methods outlined in Section 3.3. Out of the set of all possible substitutions for each question, we randomly picked atmost 2 questions to add to the training set.

We also test the performance of the model for **Multiple Word Substitutions** in the same sentence. For example, "What is the man wearing a tie pointing to?" gets perturbed to "What is the person wearing a neck-wear pointing to?". To achieve this, a profile of possible substitutions (using methods outlined in Section 3.3) is created for each token. A substitution is picked from the profile with varying probability. This is done since multiple substitutions for each relevant token with a probability p = 1.0 resulted in noise. Out of the set of distinct questions created, we randomly pick atmost 2 questions for each original question and add them to the training set.

Until now, we limit the number of augmentations generated per question to atmost 2 new questions. In this experiment, we do not restrict the number of single word substitutions performed. We refer to this experiment as **"Unrestricted Single Word Substitution"**. On an average, the number of questions per image in the augmented dataset was 33.73.

Our final experiment for Language Only Augmentation uses ConceptNet (Speer and Havasi, 2012) as a knowledgebase. ConceptNet defines nearly thirty kinds of semantic relations, most of which are not included in WordNet. We use semantic relations of the types Synonyms, Textual entailment properties (such as "HasPrerequisite","HasSubEvent"), "Used For", "Causes", "MadeOf" to generate new questions based on Hand Written Templates. We perform Single Word Substitutions and restrict the number of new augmentations to 2 new questions. We filter noisy relations based on ConceptNet's confidence score for a particular relation and whether a Surface Text in English existed or not.

The results for the experiments described above are shown in Figure 1. We observe that Unre-

Table 4: Results on VQA v2 Validation dataset stratified by Question Type. Note that only selected Question Types are shown below.

Question Type	Baseline	Object Attributes	Visual Genome	Language Only	Scene Identification	Cumulative
Do you	74.38	**76.55**	73.74	73.25	**74.44**	73.30
How many people are in	41.41	41.00	36.51	**45.14**	**42.39**	41.10
How many	42.88	40.12	38.09	**43.31**	**43.02**	40.67
Is the man	73.66	**73.87**	73.11	**74.20**	73.43	**74.46**
Is the person	75.20	**75.97**	**75.73**	**75.96**	74.33	**77.20**
Is there a	72.10	**72.57**	**72.25**	**73.23**	**72.37**	**73.50**
What color	64.89	63.63	62.18	**65.72**	**65.40**	63.31
What is the color of the	75.70	75.32	71.67	**75.99**	75.29	75.55
What is	38.47	36.68	36.16	**38.90**	38.44	37.51
What is the man	57.99	**58.42**	57.15	**59.29**	**58.00**	**58.44**
What are	53.20	**53.92**	51.63	**55.15**	**53.43**	**54.56**
What animal is	73.36	**73.99**	73.03	**73.87**	**74.37**	**74.37**
What number is	5.78	4.59	**6.10**	5.10	5.46	5.67
What room is	88.38	**89.06**	**88.81**	**88.71**	**88.54**	**88.74**
What sport is	87.86	**88.07**	**87.94**	**88.05**	**87.90**	**89.10**
Where is the	30.77	30.57	28.51	**31.01**	30.03	29.50
Which	43.88	**43.96**	43.28	43.05	**44.24**	42.56
Why	19.00	17.32	15.21	**19.19**	18.09	17.67

stricted Single Word Substitution reaches a superior performance in lesser number of Epochs. However, its performance at 20 epochs is slightly less than that for Restricted Single Word and Multiple Word Substitutions. Further, performing Multiple Word Substitutions results in a 0.09% increase in performance over Single Word Substitutions. Using ConceptNet as a knowledge base results in the slowest improvement in performance with an accuracy of 58.81% as opposed to an accuracy of 59.51% with WordNet. Possible reasons for decrease in performance could be due to noise in the ConceptNet knowledge base as well as modifications made to the vocabulary of the test set due to new answers being generated.

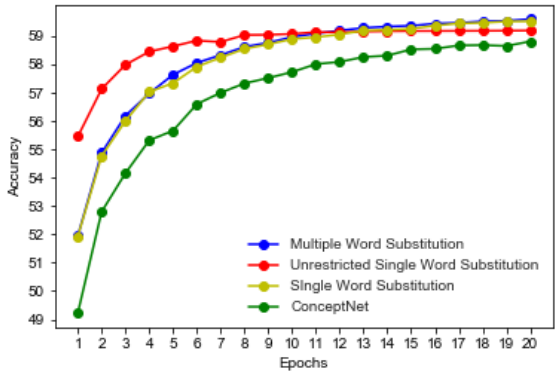

Figure 1: Comparison of Language Only Methods

Conclusion

From our experiments, we can conclude that using Language Only augmentation gives an overall increase in performance across question types. Further, using a large number of question augmentations can result in better performance with lesser epochs. The Object Attributes methods reflected an increase in performance for Object presence, recognition question types but reduced the performance for Counting and Color type questions. There might be additional complexity and reasoning required in such question types which are not captured by the augmentations. The Scene Recognition and Visual Genome method do not significantly increase the performance of any question types. The methods when cumulated together do not noticeably complement each other unless there is increase across all methods. We conclude that there is scope for better Visual Question Answering using Language Augmentation Methods.

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.".

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

GM Foody, MB McCulloch, and WB Yates. 1995. The effect of training set size and composition on artificial neural network classification. *International Journal of Remote Sensing*, 16(9):1707–1723.

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)*.

Gao Huang and Zhuang Liu. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition-Volume 1*.

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*.

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.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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.

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.

Robert Speer and Catherine Havasi. 2012. Representing general relational knowledge in conceptnet 5.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North*

American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, pages 173–180. Association for Computational Linguistics.

Kristina Toutanova and Christopher D Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*, pages 63–70. Association for Computational Linguistics.

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*.