Role of Premises in Visual Question Answering

Aroma Mahendru

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Engineering

Dhruv Batra, Chair
Devi Parikh
Bert Huang

April 10, 2017
Blacksburg, Virginia

Keywords: Machine Learning, Natural language Processing, Computer Vision

Role of Premises in Visual Question Answering

Aroma Mahendru

(ABSTRACT)

In this work, we make a simple but important observation – questions about images often contain *premises* – objects and relationships implied by the question – and that reasoning about premises can help Visual Question Answering (VQA) models respond more intelligently to irrelevant or previously unseen questions.

When presented with a question that is irrelevant to an image, state-of-the-art VQA models will still answer based purely on learned language biases, resulting in nonsensical or even misleading answers. We note that a visual question is irrelevant to an image if at least one of its premises is false (*i.e.* not depicted in the image). We leverage this observation to construct a dataset for Question Relevance Prediction and Explanation (QRPE) by searching for false premises. We train novel irrelevant question detection models and show that models that reason about premises consistently outperform models that do not.

We also find that forcing standard VQA models to reason about premises during training can lead to improvements on tasks requiring compositional reasoning.

Role of Premises in Visual Question Answering

Aroma Mahendru

(GENERAL AUDIENCE ABSTRACT)

There has been substantial recent work on the Visual Question Answering (VQA) problem in which an automated agent is tasked on answering questions about images posed in natural language. In this work, we make a simple but important observation – questions about images often contain *premises* – objects and relationships implied by the question – and that reasoning about premises can help VQA models respond more intelligently to irrelevant or previously unseen questions.

When presented with a question that is irrelevant to an image, state-of-the-art VQA models will still answer based purely on learned language biases, resulting in nonsensical or even misleading answers. We note that a visual question is irrelevant to an image if at least one of its premises is false (*i.e.* not depicted in the image). We leverage this observation to construct a dataset for Question Relevance Prediction and Explanation (QRPE) by searching for false premises. We train novel irrelevant question detection models and show that models that reason about premises consistently outperform models that do not.

We also find that forcing standard VQA models to reason about premises during training can lead to improvements on tasks requiring compositional reasoning.

Acknowledgments

First and foremost I would like to thank my advisor Dr. Dhruv Batra for his guidance and support which has been invaluable throughout this journey through graduate school. Learning from him has been an immense experience which I am sincerely grateful for.
I would also like to thank Dr. Devi Parikh for her immensely valuable feedback from time to time. Her passion and dedication are truly inspiring to me as a researcher. I am also grateful to Dr. Stefan Lee for the countless hours of discussions and the crucial help with writing on this project. This work wouldn't be the same without his guidance.
A good part of this work has been done in collaboration with Akrit and Viraj. I thank them for all their contribution and effort. This project would surely wouldn't be possible without them. I would also like to mention Neelima and Harsh for all their help with Object Proposals project. It was a great experience working with them.
I have been fortunate to be taught fun and enlightening courses by some remarkable professors. I am grateful to Dr. Dhruv Batra, Dr. Devi Parikh, Dr. Bert Huang, Dr. Scotland Leman and Dr. Mike Bowers for taking time to teach these amazing courses which were challenging and mesmerizing at the same time.
I have thoroughly enjoyed my time at CVMLP labs, especially the intense discussions during lab meetings and reading groups. I am honored to be a part of it.
Finally, I am deeply indebted to my family and close friends for all the love and support during the best and worst of times.
This work was funded in part by the following awards to DB – NSF CAREER award, ONR YIP award, ONR Grant N00014-14-1-0679, ARO YIP award, ICTAS Junior Faculty award, Google Faculty Research Award, Amazon Academic Research Award, AWS Cloud Credits for Research, and NVIDIA GPU donations. The views and conclusions contained herein are those of the author and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government or any sponsor.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The task of providing natural language answers to free-form questions about an image – *i.e.* Visual Question Answering (VQA) – has received substantial attention in the past few years [61, 5, 62, 103, 45, 97, 59, 4, 58] and has quickly become a popular problem area. Despite significant progress on VQA benchmarks [5], current models still present a number of unintelligent and problematic characteristics.

When faced with questions that are irrelevant or not applicable for an image, current 'forced choice' models will still produce an answer. For example, given an image of a dog and the question "*What color is the bird?*", standard VQA models might answer "*Red*" confidently, based only on language biases in the training set (*i.e.* an overabundance of red birds). In these cases, the predicted answers are senseless at best and misleading at worst, with either case posing serious problems for real-world applications. Like [77], we argue that practical VQA systems must be able to identify and explain irrelevant questions. For instance, a VQA model with this capability might answer "*There is no bird in the image*" when presented with this example question and image.

Premises. In this work, we make the observation that questions about images often contain *premises*, develop a premise extraction pipeline based on SPICE [3], and demonstrate how these premises can be used to address this shortcoming. Concretely, we define premises as facts implied by the language of questions, for example the question *"What brand of racket is the man holding?"* shown in Fig. 1.1 implies the existence of a man, a racket, and that the man is holding the racket. For visually grounded questions, *i.e.* questions asked about a particular image, these premises imply visual qualities of the image, including the presence of objects and their attributes and relationships.

Broadly speaking, we explore the usefulness of premises in two settings – when we know all visual questions are relevant to the images they are asked on (*e.g.* in the VQA dataset) and in real-life situations where such an assumption cannot be made (*e.g.* when asked by visually

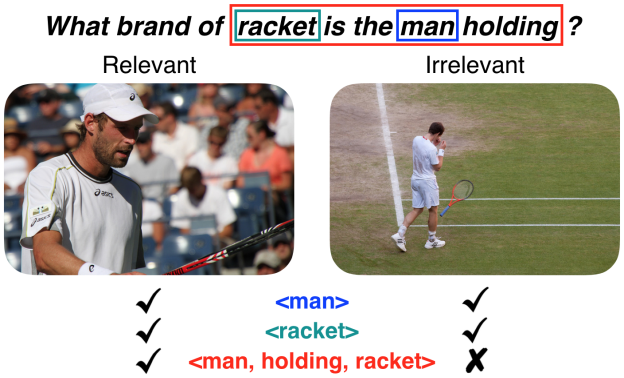


Figure 1.1: Questions asked about images often contains *'premises'* that imply visual semantics. From the above question, we can infer that a relevant image must contain a man, a racket, and that the man must be holding the racket. We extract these premises from visually grounded questions and use them to construct a new dataset and models for question relevance prediction. We also find that augmenting standard VQA training with simple premise-based questions yields improved performance on tasks requiring compositional reasoning.

impaired users). In the former case, we show that knowing that a question is relevant allows us to perform data augmentation by creating additional simple question-answer pairs using the premises of source questions. In the latter case, we show that explicitly reasoning about premises provides an effective and interpretable way of determining whether a question is relevant to an image.

Irrelevant Question Detection. We consider a question to be relevant to an image if the question's premises apply to the corresponding image *i.e.* the objects, attributes, and interactions implied by the question are depicted in the image. We refer to premises that apply for a given image to be true premises and those that do not apply as false premises. In order to train and evaluate models for this task, we curate a new irrelevant question detection dataset which we call the Question Relevance Prediction and Explanation (QRPE) dataset. QRPE is automatically curated from annotations already present in existing datasets, requiring no additional human supervision.

We collect the QRPE dataset by taking each image-question pair in the VQA dataset [5] and finding the most visually similar other image for which exactly one of the question premises is false. In this way, we collect triplets of two images and a question where the question is relevant for one image and not the other; moreover, the reason the question is irrelevant is known to be the single false premise. For context, the only other existing irrelevant question detection dataset [77] collected irrelevant question-image pairs by human verification of random pairs. In comparison, QRPE is substantially larger, balanced between irrelevant and relevant examples, and presents a considerably more difficult task due to the

closeness of the image pairs both visually and with respect to question premises. We train novel models for irrelevant question detection on the QRPE dataset and compare to existing methods. In these experiments, we show that models that explicitly reason about question premises consistently outperform baseline models that do not.

Data Augmentation. Finally, we also introduce an approach to generate simple, templated question-answer pairs about elementary concepts from premises of complex training questions. In initial experiments, we show that adding these simple question-answer pairs to VQA training data can improves performance on tasks requiring compositional reasoning. These simple questions improve training by bringing implicit training concepts "to the surface", *i.e.* introducing direct supervision of important implicit concepts by transforming them to simple training pairs.

1.1 Publications

1. Chapters 1 - 7 describe a recently submitted work:
   A. Mahendru, V. Prabhu, A. Mohapatra, D. Batra, and S. Lee. The Promise of Premise: Harnessing Question Premises in Visual Question Answering. Submitted to *EMNLP*, 2017.
2. Appendix A talks about a previous publication:
   N. Chavali, H. Agrawal, A. Mahendru, and D. Batra. Object-Proposal Evaluation Protocol is 'Gameable'. In *CVPR*, 2016.

Chapter 2

Related Work

Visual Question Answering: VQA has been the subject of a great deal of recent research attention [5, 29, 60]. Simple models like representing questions as bags of words (BoW+I [5]), or encoding the question using a recurrent neural network and train a simple classifier on the encoded question and image (Deeper LSTM [5]) have shown promise. Current top performing approaches include Neural Module Network (NMN [4]) , Hierarchical Co-Attention (HieCoAtt [59]) and Multi Model Bilinear Pooling (MCB [28]). These models use principles like explicit compositional modules, soft hierarchical co-attention and MCB kernel for pooling multimodel features respectively.

Question Relevance: Most related to our work is [77], which introduced the task of irrelevant question detection for VQA. To evaluate on this task, they created the Visual True and False Question (VTFQ) dataset by pairing VQA questions with random VQA images and having human annotators verify whether or not the question was relevant. As a result, many of the irrelevant image-question pairs exhibit a complete mismatch of image and question content. Our Question Relevance Prediction and Explanation (QRPE) dataset on the other hand is collected such that irrelevant images for each question closely resemble the source image both visually and semantically. We also provide premise-level annotations which can be used to develop models that not only decide whether a question is relevant, but also provide explanations for *why* that is the case.

Semantic Tuple Extraction: Extracting structured facts in the form of semantic tuples from text is a well studied problem [84, 3, 21]; however, recent work has begun extending these techniques to visual domains [99, 41]. Additionally, the Visual Genome [49] dataset contains dense image annotations for objects and their attributes and relationships. However, we are the first to consider these facts to reason about question relevancy and compositionality in VQA.

Textual Entailment: The task of determining whether a hypothesis sentence is true, false, or neither based on some corpus is known as textual entailment and has seen substantial work in recent years [53, 10, 92, 64] including entailment based question answering systems [82, 83]. Generating premises can be viewed as a similar task, where given a corpus (a question in our case), the goal is to extract as many statements implied by the corpus as possible rather than verify a particular statement.

Compositionality: Recently, some effort has been seen in making VQA and other joint vision and language models like image captioning more compositional. This is still an open problem which has largely been approached from a model standpoint. For example, [36] integrate data and transfer knowledge between semantically related concepts, to improve upon current deep image captioning models. Similarly, [4] construct and learn neural module networks, which composes collections of jointly-trained neural modules into deep network for VQA. These approaches are different from our work since we prepose to improve compositionality via data augmentation.

Interpretable VQA Systems: Designing interpretable deep learning systems that provide rationales for their decisions is a topic that has received much attention of late. In ([71, 95], VQA models have been presented which can support their answers with explanations. However, there has not been work on generating explanations for false premise detection such as the approach proposed in this work.

Question Premise: To the best of our knowledge no other work has used Question Premises for improving VQA models. [39] however, comes close in this regard. This work proposes use of *basic questions* as a means to improve performance of VQA model.

However, basic questions are less complex questions that the main question can be dissociated into as opposed to premise questions, which are generated from implied facts or premises in the question. They also formulate basic question generation as a learning problem *i.e.* train a model on a collected dataset, while we use a templated rule based pipeline for premise question generation. Our work is also different in the aspect that we use premise questions to improve compositionality, question relevance prediction and explanation. [39] uses basic questions to improve accuracy on the VQA task by adding basic question features along with main question and image features to the classifier.

Chapter 3

Premise Extraction

We now formalize the concept of premises and explain how premises can be extracted from questions. As discussed in Chapter 1, we define question premises as facts implied about an image from a question asked about it, which we represent as tuples. Returning to our running example question *"What brand of racket is the man holding?"*, we can express these premises as the tuples *'⟨man⟩'*, *'⟨racket⟩'*, and *'⟨man, holding, racket⟩'* respectively.

In this work, we categorize these tuples into three groups based on their complexity. First-order premises represent the presence of objects (*'⟨man⟩'*, *'⟨cat⟩'*, *'⟨sky⟩'*), second-order premises capture the attributes of objects (*'⟨man, tall⟩'*, *'⟨car, moving⟩'*), and third-order premises contain interactions between objects (*e.g.* *'⟨man, kicking, ball⟩'*, *'⟨cat, above, car⟩'*).

Premise Extraction: To extract premises from questions, we use the semantic tuple extraction pipeline used in the SPICE metric [3]. Originally defined as a metric for image captioning, SPICE transforms a sentence into a scene graph using the Stanford Scene Graph Parser [84] and then extracts semantic tuples from this representation. Fig. 3.1 shows this process for a sample question. The question is represented as a graph of objects, attributes, and relationships from which first, second, and third order premises are extracted respectively. As this pipeline was originally designed for descriptive captions rather than questions, we found a number of minor modifications helpful in extracting quality question premises, including disabling pronoun resolution, verb lemmatization and METEOR-based Synset matching. We will release our premise extraction code publicly to encourage reproducibility.

While this extraction process typically produces high quality premise tuples, there are some sources of noise which must be filtered out. The SPICE process occasionally produces duplicate nodes or object nodes not linked to nouns in the question, which we filter out. We also remove premises containing words like photo, image, *etc.* that refer to the image rather than its content.

A more nuanced source of erroneous premises comes from the ambiguity in existential ques-

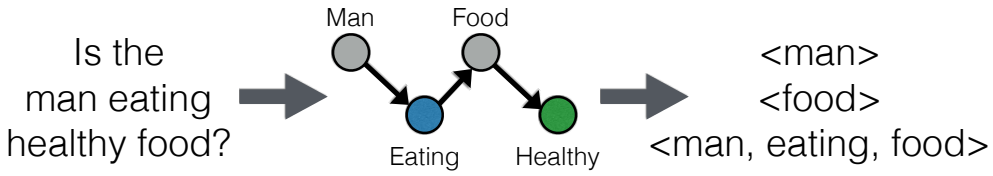


Figure 3.1: **Premise Extraction Pipeline.** Objects (gray), attributes (green), and relations (blue) scene graph nodes are converted into 1st, 2nd, and 3rd order premises respectively.

tions, *i.e.* those about the existence of certain image content. For example, while the question *"Is the little girl moving?"* contains the premise '*⟨girl, little⟩*', it is unclear without the answer whether '*⟨girl, moving⟩*' is also a premise. Similarly, for the question *"How many giraffes are in the image?"*, '*⟨giraffe, many⟩*' cannot be considered a premise as there may be 0 giraffes in the image. To avoid introducing false premises, we filter out existential and counting questions which are answered negatively (either "no" or "0") if ground truth is available. Otherwise we threshold SPICE similarity between generated tuples and source questions to avoid repeating ambgious premises.

Chapter 4

Question Relevance Prediction and Explanation Dataset

As discussed in Chapter 1, modern VQA models fail to differentiate between relevant and irrelevant questions, answering either with confidence. This behavior is detrimental to the real world application of VQA systems. In this chapter, we curate a new dataset for question relevance in VQA which we call the Question Relevance Prediction and Explanation (QRPE) dataset. We plan to release QRPE publicly to help future efforts.

In order to train and evaluate models for irrelevant question detection, we would like to create a dataset of triplets (I^+, Q, I^-) comprised of a natural language question Q, an image I^+ for which Q is relevant, and an image I^- for which Q is irrelevant. While it is not required to collect both a relevant and irrelevant image for each question, we argue that doing so is a simple way to balance the dataset and it ensures that biases against rarer questions (which would be irrelevant for most images) cannot be exploited to inflate performance.

We base our dataset on the existing VQA corpus [5], taking the human-generated (and therefore relevant) image-question pairs from VQA as I^+ and Q. As previously discussed, we can define the relevancy of a question in terms of the grounding of its premises within an image, so we extract premises from each question Q and must find a suitable irrelevant image I^-. However, there are certainly many images for which one or more of Q's premises are false. One design decision is then how to select I^- from this set.

To ensure our dataset is as realistic and challenging as possible, we consider irrelevant images which have only a single false question premise under Q. For example the question *"Is the big red dog old?"* could be matched with an image containing a big, white dog or a small red dog, but not a small white dog. In this way, we ensure that image content is semantically appropriate for the question topic but not quite relevant. Additionally, this provides each irrelevant image with an explanation for why the question does not apply.

Furthermore, we sort this subset of irrelevant image by their visual distance to the source

Figure 4.1: **Some Examples from QRPE Dataset.** For a given question Q and a relevant image I^+, we find an irrelevant image I^- for which exactly one premise of the question is false. If there are multiple such candidates, we select the candidate most visually most similar to I^+. As can be seen from these examples, the QRPE dataset is very challenging, with only minor visual and semantic differences separating the relevant and irrelevant images.

image I^+ based on image encodings from a VGGNet [86] pretrained on ImageNet [80]. This ensures that the relevant and irrelevant images are visually similar and act as difficult examples.

A major difficulty with our proposed data collection process is how to verify whether a premise if true or false for any given image in order to identify irrelevant images. We detail dataset construction and our approach for this problem in the following section.

4.1 Dataset Construction

We curate our QRPE dataset automatically from existing annotations in COCO and Visual Genome. For first order premises (*i.e.* existential premises), we consider only the 80 classes present in COCO [56]. As VQA and COCO share the same images, we can easily determine if a first order premise is true or false for a candidate irrelevant image simply by checking for the absence of the appropriate class annotation.

For second order premises (*i.e.* attributed objects), we rely on Visual Genome [49] annotations for object and attribute labels. Unlike in COCO, the lack of a particular object label in an image for Visual Genome does not necessarily indicate that the object is not present, due both to annotation noise and the use of multiple synonyms for objects by human labelers. As a consequence, we restrict the set of candidate irrelevant images to those which contain a matching object to the question premise but a different attribute. Without further restriction, the selected irrelevant attributes do not tend to be mutually exclusive with the

source attribute (*i.e.* matching *'⟨dog, old⟩'* and *'⟨dog, red⟩'*). To correct this and ensure a false premise, we further restrict the set to attributes which are antonyms (*e.g.* *'⟨young⟩'* for source attribute *'⟨old⟩'*) or taxonomic sister terms (*e.g.* *'⟨green⟩'* for source attribute *'⟨red⟩'*) of the original premise attribute. We also experimented with third order premises; however, the lack of a corresponding sense of mutual exclusion for verbs and the sparsity of ⟨object, relationship, object⟩ premises made finding non-trivial irrelevant images difficult.

To recap, our data collection approach is then to take each image-question pair in the VQA dataset and extract its first and second order question premises. For each premise, we find all images which lack only this premise and rank them by their visual distance. The closest of these is kept as the irrelevant image for each image-question pair.

4.2 Exploring the Dataset

Fig. 4.1 shows sample (I^+, Q, I^-) triplets from our dataset along with the falsified premise that makes I^- irrelevant for Q. These examples illustrate the difficulty of our dataset. The images in the second column differ only because a red firetruck is not a red bus and the final column are differentiated only by the number of giraffe. Both of these are fine details of the image content.

The QRPE dataset contains 102,432 (I^+, Q, I^-) triplets generated from as many premises. In total, it contains 2961 unique premises and 96,812 unique questions. Among the 102,432 premises, 11,065 are second-order, attributed object premises while the remaining 91,367 are first-order object/scene premises. We divide our dataset into two parts – a training set with 68,037 premises that is generated from the VQA training set, and a validation set with 34,395 premises, generated from the VQA validation set.

4.3 Comparison to VTFQ

We contrast our approach to the VTFQ dataset of [77]. As discussed prior, VTFQ was collected by selecting a random question and image from the VQA set and asking human annotators to report if the question was relevant. This approach results in irrelevant image-question pairs that are unambiguously unrelated, with the visual content of the image having nothing at all to do with the question or its source image from VQA.

In Fig. 4.2, we present a quantitative and qualitative comparison of the two datasets. For the sake of comparison, we generate (I^+, Q, I^-) triplets from VTFQ by find the nearest neighbor question Q^{nn} in the VQA dataset to Q for each (Q, I^-) pair in VTFQ. We then select the image on which Q^{nn} was asked as I^+. We plot the Euclidean distance between the fc7 features of each (I^+, I^-) pair in both datasets. As shown on the left side of Fig. 4.1, we

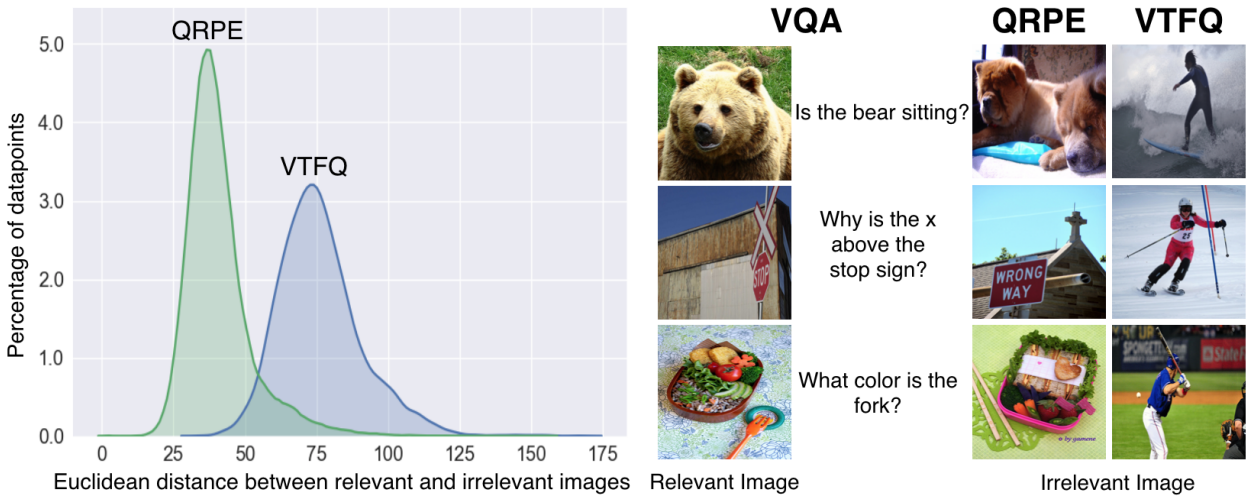


Figure 4.2: **A comparison of the QRPE and VTFQ Datasets.** On the left, we plot the Euclidean distance between VGGNet-fc7 features extracted from each relevant-irrelevant image pair for each dataset. Note that VTFQ has significantly higher visual distances. On the right, we show some qualitative examples of irrelevant images for questions that occur in both datasets. VTFQ images are significantly less related to the source image and question than in our dataset.

find that the mean distance in the VTFQ dataset is nearly twice that of our QRPE dataset, indicating that irrelevant images in VTFQ are less visually related to source images.

On the right side of Fig. 4.2, we also provide qualitative examples of questions that occur in both datasets. The example on the last row is perhaps most striking. The source question is asking the color of a fork and the relevant image shows an overhead view of a meal with an orange fork set nearby. The irrelevant image in QRPE is a similar image of food, but with chopsticks! Conversely, the image from VTFQ is a man playing baseball.

Chapter 5

Question Relevance Detection

In this chapter, we introduce a simple baseline for irrelevant question detection on the QRPE dataset and demonstrate that explicitly reasoning about premises improves performance for both our new model and existing methods. More formally, we consider the binary classification task of predicting if a question Q_i from an image-question pair (I_i, Q_i) is relevant to image I_i.

A Simple Premise-Aware Model. Like the standard VQA task, question relevance detection also requires making a prediction based on an encoded image and question. With this in mind, we begin with a straight-forward approach based on the Deeper LSTM VQA model architecture of [5]. This model encodes the image I via a VGGNet and the question Q with an LSTM over one-hot word encodings. The concatenation of these embeddings are input to a multi-layer perceptron. We fine-tune this model for the binary question relevance detection task starting from a model pretrained on the VQA task. We denote this model as `VQA-Bin`.

We extend the `VQA-Bin` model to explicitly reason about premises. We extract first and second order premises from the question Q and encode them as two concatenated one-hot vectors. We add an additional LSTM to encode the premises and concatenate this added feature to the image and question feature. We refer to this premise-aware model as `VQA-Bin-Premise`.

Existing Methods. We compare our approaches with the best performing model of [77]. This model (which we call `QC-Sim`) uses NeuralTalk2 [44] trained on the MS COCO dataset [56] to generate a caption for each image and trains a multilayer perceptron to learn a similarity between LSTM embeddings (with shared weights) the question and generated caption (encoded as word2vec [65] representations). We consider two additional versions of this approach that consider premise-caption similarity (`PC-Sim`) and question-premise-

Models	Accuracy
VQA-Bin	65.9%
VQA-Bin-Prem	66.3%
QC-Sim	73.6%
PC-Sim	74.1%
QPC-Sim	**74.5%**

Table 5.1: Accuracy of Question Relevance models on the QRPE test set. We find that premise-aware models consistently outperform alternative models.

caption similarities (`QPC-Sim`). We note that the caption similarity based methods makes use of significant outside data through NeuralTalk2 captioning and word2vec embeddings.

Results. We train each model on the QRPE train split and report results on the test set in Table 5.1. We find that the addition of extracted premise representations consistently improves peformance of base models. This is especially interesting given that the models *already* have access to the question from which the premises were extracted. This result seems to imply there is value in explicitly isolating premises from sentence grammar.

We find that the caption similarity models significantly outperform our proposed VQA style models, with the overal best performing approach being premise augmented `QPC-Sim` model.

5.1 Question Relevance Explanation

Figure 5.1: **Question relevance explanation:** We provide selected examples of predictions from the False Premise Detection model (`FPD`) on the QRPE test set. Reasoning about premises presents the opportunity to produce natural language statements indicating *why* a question is irrelevant to an image, by pointing to the premise that is invalid.

In addition to identifying whether a question is irrelevant to an image, being able to indicate *why* carries significant real-world utility. From an interpretability perspective, reporting which premise is false is more informative than simply answering the question in the negative, as it can help to correct the questioner's misconception regarding the scene. We propose to generate such explanations by identifying the particular question premise(s) that do not apply to an image.

By construction, irrelevant images in the QRPE dataset are picked on the basis of negating a single premise – we now use our dataset to train models to detect false premises, and use the premises classified as irrelevant to generate templated natural language explanations.

Fig. 5.1 illustrates the task setup for false premise detection. Given a question-image pair, say *"Is the dog chasing the sheep?"*, the objective is to identify which (if any) question premises are not grounded in the image, in this case both *⟨dog⟩* and *⟨sheep⟩*. Alternatively, for the question *"What color are the four cones next to the truck?"*, both premises *⟨truck⟩* and *⟨cones, four⟩* are true premises grounded in the image.

For this task we train a binary classifier similar to which uses a one-hot encoding of premises (generates from the vocabulary of all premise words) and features from the last hidden layer of VGGNet [86] to represent the image. We concatenate these features and feed them into a multilayer perceptron which predicts whether the premise is grounded in the image or not. We trained our false premise detection model (`FPD`) model on all premises in the QRPE dataset.

Our `FPD` model achieves an accuracy of 60.8% on the QRPE dataset. In Fig. 5.1, we present qualitative results of our premise classification and explanation pipeline. For the question *"Is the dog chasing the sheep?"*, the model correctly recognizes 'dog' and 'sheep' as false premises, and we generate statements in natural language indicating the same. Thus, determining question relevance by reasoning about each premise presents the opportunity to generate simple explanations that can provide valuable feedback to the questioner, and help improve model trust.

Chapter 6

Premise-Based Data Augmentation for VQA

In this chapter, we develop a premise-based data augmentation scheme for VQA that generates simple, templated questions based on premises present in complex visually-grounded questions from the VQA training set. As these question-image pairs were collected from sighted humans instructed to ask questions about a given an image, we assume that the questions (and the premises they imply) are grounded in the objects and relationships depicted in the corresponding image. We discuss question generation pipeline and various data augmentation experiments performed with these premise questions in Sections 6.1, 6.2 respectively.

6.1 Question Generation

Using the pipeline presented in Chapter 3, we extract premises from questions in the VQA dataset and apply a simple templated question generation strategy to transform premises into question and answer pairs. This process transforms implicit premise concepts which previously had to be learned as part of understanding more complex questions, into simple, explicit training examples that can be directly supervised. The generated premise questions cover a large number of objects, attributes, and relations, and as a result, including them in the VQA training set greatly expands and alters the answer space distribution. We designed specific templates for each type of premise.

First order facts like ⟨man⟩, ⟨bus⟩ lead to existential questions like "Is there a man?", "Is there a bus?" and so on. Second order facts can generate two kinds of questions depending on whether the second element is an action or an attribute. For example, ⟨man, walking⟩ would become "What is the man doing?" while ⟨car, red⟩ would become "What is the color of the car?". In general, questions generated from third order facts look like "Is the man

Figure 6.1: **Question generation** For every source question, premise tuples are extracted and then used to generate premise questions using a rule-based NLP pipeline.

holding the racket?", and "What is the cat on top of?" for ⟨man, holding, racket⟩ and ⟨cat, on top of, box⟩, respectively. However, third order facts are slightly more complicated and many kinds of questions can be generated from them depending on the types of elements present.

Question generation also involves minor pre-processing and post-processing. In the former we remove erroneous premises from the set while in the latter we remove generated questions which are linguistically ambiguous. We also run SPICE on the generated questions using the source questions as references to eliminate generated questions that are duplicates of the source questions. A random selection of premise questions generated from the VQA dataset can be seen in Fig. 6.2. The answer type distribution of generated premise questions can be seen in Table 6.1. We find that generated premise questions are twice in number as compared to source questions. It can also be seen that very few 'Number' questions are extracted. This is because people generally do not ask questions about multiple number of objects at the same time. By design, we generate only 'Yes' questions and zero 'No' questions. The reason for that is twofold – first, we only generate premise questions from true premises, and second, first order premises are the most frequent premises in source questions (first order premises generate 'Yes' questions).

Training Data	Other	Number	Yes	No	Total
Source	123,817	29,698	57217	35842	246,574
Premise	137,483	1,850	387,941	0	527,274

Table 6.1: Answer type distribution of source and premise questions on the Compositional VQA train set.

What will happen when the finger pushes the button?

What is the finger pushing? Button

Is there a finger in the image? Yes

Is there a button in the image? Yes

Has someone already eaten off the plate?

What is the someone eating off? Plate

Is there a someone in the image? Yes

Is there a plate in the image? Yes

What is the item called that the cat is looking at?

Is there a cat in the image? Yes

Is there an item in the image? Yes

What is the child sitting on?

What is the child doing? Sitting

Is there a child in the image? Yes

What player number is about to swing at the ball?

Is there a player number? Yes

Is there a number in the image? Yes

Is there a ball in the image? Yes

What is the man carrying in his left hand?

Who is carrying in the hand? Man

What is the man carrying in? Hand

Figure 6.2: Sample generated premise questions from source questions. Source questions are in bold. Ground-truth answers are extracted using the premise tuples.

6.2 Data Augmentation

Using the premise questions generated by employing the approach described in Section 6.1, we train VQA models on the augmented training set with various augmentation strategies based on different subsets of the premise questions. These ablations are described as follows:

1. **None:** No premise questions added to the training set.
2. **All:** Adding all the generated premise questions along with source questions to the training set.
3. **Only-Binary:** Only Binary (Questions with answers "Yes/No") premise questions added along with the source questions.
4. **No-Other:** All questions except the type Other premise questions (answers outside of Binary and Number answers) added to the training set.
5. **No-Binary:** All questions except Binary premise question types added to the training set.
6. **Comm-Other:** All Binary premise questions added. From Other and Number premise question types, selected premise questions are added whose answers lie in the source question answer space.
7. **Top1k-A:** All Binary premise questions added. From 'Others', selected premise questions are added whose answers are amongst the most top1000 VQA source responses.

Note that evaluation is performed on standard validation set containing no premise questions. We evaluate performance of VQA models trained with these various ablations on both -

Standard VQA split as well as the Compostitional VQA split [1].

6.2.1 Results and Analysis

	Augmentation	Overall	Other	Number	Yes/No
Standard	None	54.23	40.34	33.27	79.82
	All	53.74	39.28	**33.38**	79.89
	Top-1k-A	**54.47**	**40.56**	33.24	**80.19**
Comp.	None	46.69	31.92	29.73	70.49
	All	47.63	31.97	**30.77**	**72.52**
	Top-1k-A	**47.85**	**32.58**	30.59	72.38

Table 6.2: Accuracy on the standard and compositional VQA validation sets for different augmentation strategies.

Standard VQA Split. We evaluate the augmentation settings described above with the Deeper LSTM/CNN VQA model by [57] on the Standard VQA split and show the results in the top half of Table 6.3. We find minor improvements of 0.34% using when restricting premise questions to match VQA answers (Top-1k-A).

Data Ablation	Overall	Other	Number	Yes/No
None	46.69	31.92	29.73	70.49
All	47.63	31.97	**30.77**	**72.52**
Only-Binary	47.25	32.45	29.65	71.30
No-Other	47.33	32.47	29.85	71.42
No-Binary	46.76	31.69	29.39	71.09
Comm-Other	47.53	32.41	28.88	72.33
Top1k-A	**47.85**	**32.58**	30.59	72.38

Table 6.3: Performance of DeeperLSTM [5] on Compositional VQA test split with different augmentations.

Compositional VQA Split. We also evaluate on a custom compositional VQA dataset split [1] that is specifically designed to test a model's ability to generalize to unseen/rarely seen combinations of concepts at test time. The bottom half of Table 6.3 shows results on this split. With our best ablation we observe a 1.16% gain in VQA accuracy over no augmentation. In this setting, explicitly reasoning about objects and attributes seen in the questions

VQA Model	Baseline	With Premises
DeeperLSTM[5]	46.69	47.85
HieCoAtt[59]	50.17	49.98
NMN[4]	49.05	48.43
MCB[28]	50.13	50.57

Table 6.4: Accuracy of different VQA models on the Compositional VQA test split using Top1k-A augmentation.

helps the model to disentangle objects from their most common characteristics. Some qualitative examples where an augmented model performs better than a non-augmented model are shown in Fig. 6.3.

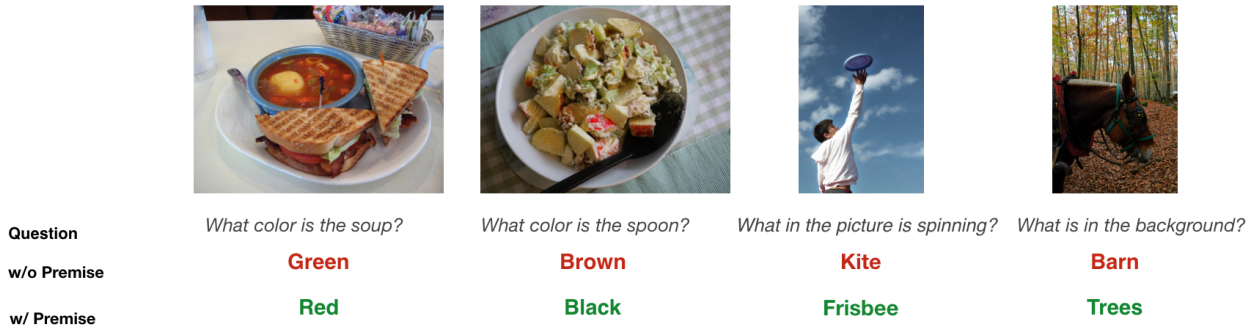


Figure 6.3: Some interesting examples of how augmentation helps the DeeperLSTM model [5] on the compositional VQA split.

We also train Deeper LSTM model on various other ablations of premise question and evaluate their performance. We observe that most of the ablations (Table 6.3) using premises show some improvement in performance over the model trained with no premise questions at all. Finally, we try replacing Deeper LSTM model with other VQA models to see if the perfomance boost is replicated. We observe that though data augmentation using premise questions help some models, there are other models for which that's not the case (Table 6.4).

Chapter 7

Conclusion and Future Work

In this work, we made the simple observation that questions about images often contain premises implied by the question and showed that reasoning about premises can help VQA models respond more intelligently to irrelevant or previously unseen questions.

We develop a system for automatically extracting these question premises. Using extracted premises, we automatically created a novel dataset for Question Relevance Prediction and Explanation (QRPE) which consists of 102,432 question, relevant image, and irrelevant image triplets. We also train novel question relevance prediction models and show that models that take advantage of premise information outperform models that do not. Furthemore, we demonstrated that questions generated from premises may be an effective data augmentation techniques for VQA tasks that require compositional reasoning.

7.1 Future Work

Integrating Question Relevance Prediction and Explanation models with existing VQA systems would form a natural extension to VQA models. In this setting, the Relevance Prediction model would determine the applicability of a question to an image, and select an appropriate path of action. If the question is classified as relevant, the VQA model would generate a prediction; otherwise, a Question Relevance explanation model would provide a natural language sentence indicating which premise(s) are not valid for the image.

Premises can also aid users in deciding when to place trust in the predictions from a VQA model. If a VQA model answers a question correctly, but is unable to correctly answer the simple premise questions (as described in Chapter 6), it would indicate that the model does not understand the underlying concepts in the question. However, being able to answer both source and premise questions correctly would suggest that the model understands underlying concepts, and thus the prediction can be trusted.

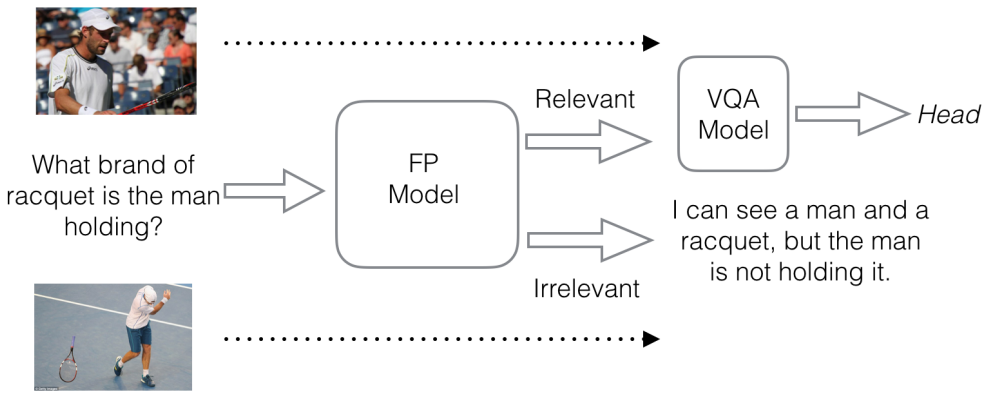


Figure 7.1: A complete VQA system that can additionally determine and explain the applicability of a question to an image.

Such applications of question premises can help VQA systems to take a step in the direction of moving beyond academic settings to real-world environments.

Appendix A

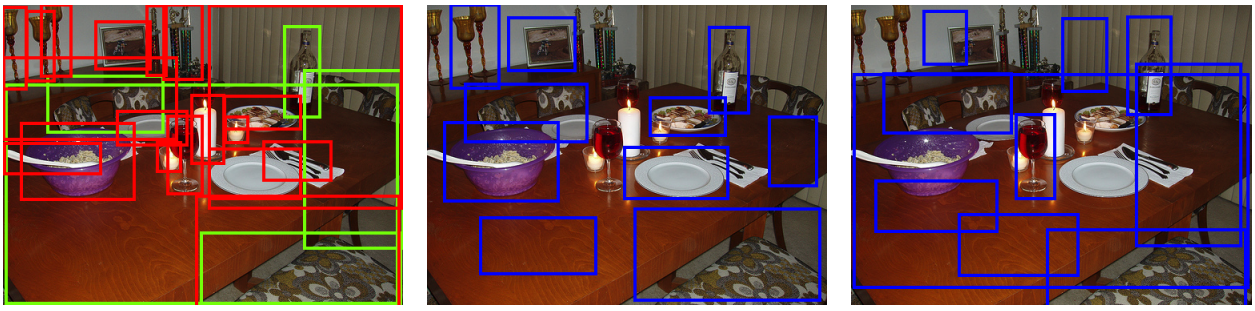
Object Proposals

Object proposals have quickly become the de-facto pre-processing step in a number of vision pipelines (for object detection, object discovery, and other tasks). Their performance is usually evaluated on partially annotated datasets. In this work, we argue that the choice of using a partially annotated dataset for evaluation of object proposals is problematic – as we demonstrate via a thought experiment, the evaluation protocol is 'gameable', in the sense that progress under this protocol does not necessarily correspond to a "better" category independent object proposal algorithm.

To alleviate this problem, we: (1) Introduce a nearly-fully annotated version of PASCAL VOC dataset, which serves as a test-bed to check if object proposal techniques are overfitting to a particular list of categories. (2) Perform an exhaustive evaluation of object proposal methods on our introduced nearly-fully annotated PASCAL dataset and perform cross-dataset generalization experiments; and (3) Introduce a diagnostic experiment to detect the *bias capacity* in an object proposal algorithm. This tool circumvents the need to collect a densely annotated dataset, which can be expensive and cumbersome to collect. Finally, we plan to release an easy-to-use toolbox which combines various publicly available implementations of object proposal algorithms which standardizes the proposal generation and evaluation so that new methods can be added and evaluated on different datasets. We hope that the results presented in the work will motivate the community to test the category independence of various object proposal methods by carefully choosing the evaluation protocol.

A.1 Introduction

In the last few years, the Computer Vision community has witnessed the emergence of a new class of techniques called *Object Proposal* algorithms [104, 22, 7, 2, 75, 63, 76, 91, 40, 14, 46].



(a) (Green) Annotated, (Red) Unannotated

(b) Method 1 with recall 0.6 on the image

(c) Method 2 with recall 1 on the image

Figure A.1: (a) shows PASCAL annotations natively present in the dataset in green. Other objects that are not annotated but present in the image are shown in red; (b) shows Method 1 and (c) shows Method 2. Method 1 visually seems to recall more categories such as plates, glasses, *etc*. that Method 2 missed. Despite that, the computed recall for Method 2 is higher because it recalled all instances of PASCAL categories that were present in the ground truth. Note that the number of proposals generated by both methods is equal in this figure.

Object proposals are a set of candidate regions or bounding boxes in an image that may potentially contain an object.

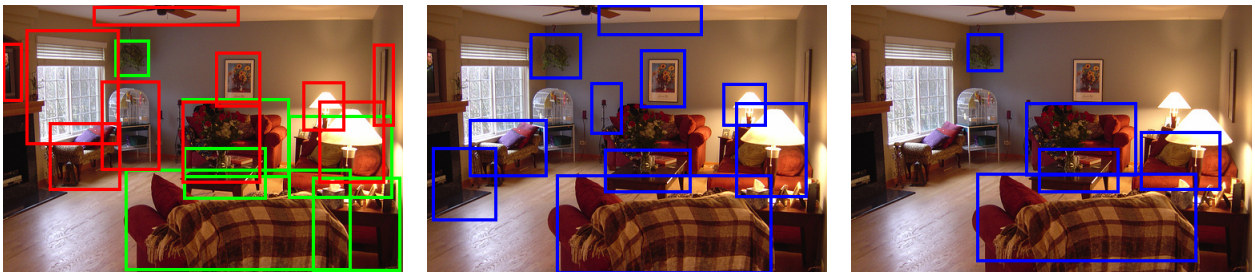
Object proposal algorithms have quickly become the de-facto pre-processing step in a number of vision pipelines – object detection[31, 33, 89, 96, 17, 23, 51, 90, 102, 34], segmentation[12, 6, 11, 18, 101], object discovery[19, 15, 42, 79], weakly supervised learning of object-object interactions[16, 74], content aware media re-targeting[87], action recognition in still images[85] and visual tracking[94, 52]. Of all these tasks, object proposals have been particularly successful in object detection systems. For example, *nearly all top-performing entries*[88, 54, 70, 33] in the ImageNet Detection Challenge 2014 [81] used object proposals. They are preferred over the formerly used sliding window paradigm due to their computational efficiency. Objects present in an image may vary in location, size, and aspect ratio. Performing an exhaustive search over such a high dimensional space is difficult. By using object proposals, computational effort can be focused on a small number of candidate windows.

The focus of this work is the protocol used for evaluating object proposals. Let us begin by asking – *what is the purpose of an object proposal algorithm?*

In early works [2, 22, 63], the emphasis was on *category independent object proposals*, where the goal is to identify instances of *all* objects in the image irrespective of their category. While it can be tricky to precisely define what an "object" is[1], these early works presented cross-category evaluations to establish and measure category independence.

More recently, object proposals are increasingly viewed as *detection proposals* [46, 91, 104, 47]

[1]Most category independent object proposal methods define an object as "stand-alone thing with a well-defined closed-boundary". For "thing" *vs*. "stuff" discussion, see [35].



(a) (Green) Annotated, (Red) Unannotated

(b) Method 1 with recall 0.5 on the image

(c) Method 2 with recall 0.83 on the image

Figure A.2: (a) shows PASCAL annotations natively present in the dataset in green. Other objects that are not annotated but present in the image are shown in red; (b) shows Method 1 and (c) shows Method 2. Method 1 visually seems to recall more categories such as lamps, picture, *etc*. that Method 2 missed. Clearly the recall for Method 1 *should* be higher. However, the calculated recall for Method 2 is significantly higher, which is counter-intuitive. This is because Method 2 recalls more PASCAL category objects.

where the goal is to improve the object detection pipeline, focusing on a chosen set of object classes (*e.g.* ˜20 PASCAL categories). In fact, many modern proposal methods are learning-based[14, 46, 40, 47, 43, 50, 30, 72] where the definition of an "object" is the set of annotated classes in the dataset. This increasingly blurs the boundary between a proposal algorithm and a detector.

Notice that the former definition has an emphasis on object discovery[19, 15, 79], while the latter definition emphasises on the ultimate performance of a detection pipeline. Surprisingly, despite the two different goals of 'object proposal,' there exists only a single evaluation protocol:

1. Generate proposals on a dataset: The most commonly used dataset for evaluation today is the PASCAL VOC [24] detection set. Note that this is a *partially annotated* dataset where only the 20 PASCAL category instances are annotated.

2. Measure the performance of the generated proposals: typically in terms of 'recall' of the annotated instances. Commonly used metrics are described in Section A.3.

The central thesis of this work is that the current evaluation protocol for object proposal methods is suitable for object detection pipeline but is a *'gameable' and misleading protocol* for category independent tasks. By evaluating only on a specific set of object categories, we fail to capture the performance of the proposal algorithms on *all the remaining object categories that are present in the test set, but not annotated in the ground truth.*

Figs. A.1, A.2 illustrate this idea on images from PASCAL VOC 2010. Column (a) shows the ground-truth object annotations (in green, the annotations natively present in the dataset

for the 20 PASCAL categories –'chairs', 'tables', 'bottles', *etc.*; in red, the annotations that we added to the dataset by marking object such as 'ceiling fan', 'table lamp', 'window', *etc.* originally annotated 'background' in the dataset). Columns (b) and (c) show the outputs of two object proposal methods. Top row shows the case when both methods produce the same number of proposals; bottom row shows unequal number of proposals. We can see that proposal method in Column (b) seems to be more "complete", in the sense that it recalls or discovers a large number of instances. For instance, in the top row it detects a number of non-PASCAL categories ('plate', 'bowl', 'picture frame', *etc.*) but misses out on finding the PASCAL category 'table'. In both rows, the method in Column (c) is reported as achieving a higher recall, *even in the bottom row, when it recalls strictly fewer objects, not just different ones*. The reason is that Column (c) recalls/discovers instances of the 20 PASCAL categories, which are the only ones annotated in the dataset. Thus, Method 2 appears to be a *better* object proposal generator simply because it focuses on the annotated categories in the dataset.

While intuitive (and somewhat obvious) in hindsight, we believe this is a crucial finding because it makes the current protocol *'gameable'* or susceptible to manipulation (both intentional and unintentional) and misleading for measuring improvement in category independent object proposals.

Some might argue that if the end task is to detect a certain set of categories (20 PASCAL or 80 COCO categories) then it is enough to evaluate on them and there is no need to care about other categories which are not annotated in the dataset. We agree, but it is important to keep in mind that object detection is not the only application of object proposals. There are other tasks for which it is important for proposal methods to generate category independent proposals. For example, in semi/unsupervised object localization[19, 15, 42, 79] the goal is to identify all the objects in a given image that contains many object classes without any specific target classes. In this problem, there are no image-level annotations, an assumption of a single dominant class, or even a known number of object classes[15]. Thus, in such a setting, using a proposal method that has tuned itself to 20 PASCAL objects would not be ideal – in the worst case, we may not discover any new objects. As mentioned earlier, there are many such scenarios including learning object-object interactions[16, 74], content aware media re-targeting[87], visual tracking[52], *etc.*

To summarize, the contributions of this work are:

- We report the 'gameability' of the current object proposal evaluation protocol.
- We demonstrate this 'gameability' via a simple thought experiment where we propose a 'fraudulent' object proposal method that *significantly outperforms all existing object proposal techniques* on current metrics, but would under any no circumstances be considered a category independent proposal technique. As a side contribution of our work, we present a simple technique for producing state-of-art object proposals.
- After establishing the problem, we propose three ways of improving the current eval-

uation protocol to measure the category independence of object proposals:

1. evaluation on *fully* annotated datasets,
2. cross-dataset evaluation on *densely* annotated datasets.
3. a new evaluation metric that quantifies the *bias capacity* of proposal generators.

For the first test, we introduce a nearly-fully annotated PASCAL VOC 2010 where we annotated *all instances of all object categories* occurring in the images.

- We thoroughly evaluate existing proposal methods on this nearly-fully and two densely annotated datasets.
- We will release all code and data for experiments, and an object proposals library that allows for easy comparison of all popular object proposal techniques.

A.2 Related Work

Types of Object Proposals: Object proposals can be broadly categorized into two categories:

- **Window scoring**: In these methods, the space of all possible windows in an image is sampled to get a subset of the windows (*e.g.*, via sliding window). These windows are then scored for the presence of an object based on the image features from the windows. The algorithms that fall under this category are [2, 75, 104, 14, 13, 30].
- **Segment based**: These algorithms involve over-segmenting an image and merging the segments using some strategy. These methods include [91, 76, 7, 46, 40, 22, 63, 93, 50, 72]. The generated region proposals can be converted to bounding boxes if needed.

Beyond RGB proposals: Beyond the ones listed above, a wide variety of algorithms fall under the umbrella of 'object proposals'. For instance, [69, 27, 100, 66, 98] used spatio-temporal object proposals for action recognition, segmentation and tracking in videos. Another direction of work[32, 8, 9] explores use of RGB-D cuboid proposals in an object detection and semantic segmentation in RGB-D images. While the scope of this work is limited to proposals in RGB images, the central thesis of the work (*i.e.*, gameability of the evaluation protocol) is broadly applicable to other settings.

Evaluating Proposals: There has been a relatively limited analysis and evaluation of proposal methods or the proposal evaluation protocol. Hosang *et al.* [38] focus on evaluation of object proposal algorithms, in particular the stability of such algorithms on parameter changes and image perturbations. Their works shows that a large number of category independent proposal algorithms indeed generalize well to non-PASCAL categories, for instance

in the ImageNet 200 category detection dataset [81]. Although these findings are important (and consistent with our experiments), they are unrelated to the 'gameability' of the evaluation protocol, which is our focus. In [37], authors present an analysis of various proposal methods regarding proposal repeatability, ground truth annotation recall, and their impact on detection performance. They also introduced a new evaluation metric (Average Recall). Their argument for a new metric is the need for a better localization between generated proposals and ground truth. While this is a valid and significant concern, it is orthogonal to the'gameability' of the evaluation protocol, which to the best of our knowledge has not been previously addressed. Another recent related work perhaps is [73], which analyzes the state-of-the-art methods in segment-based object proposals, focusing on the challenges faced when going from PASCAL VOC to MS COCO. They also analyze how aligned the proposal methods are with the bias observed in MS COCO towards small objects and the center of the image and propose a method to boost their performance. Although there is a discussion about biases in datasets but it is unlike our theme, which is 'gameability' due to these biases. As stated earlier, while early papers [2, 22, 63] reported cross-dataset or cross-category generalization experiments similar to ones reported in this work, with the trend of learning-based proposal methods, these experiments and concerns seem to have fallen out of standard practice, which we show is problematic.

A.3 Evaluating Object Proposals

Before we describe our evaluation and analysis, let us first look at the object proposal evaluation protocol that is widely used today. The following two factors are involved:

1. **Evaluation Metric**: The metrics used for evaluating object proposals are all typically functions of intersection over union (IOU) (or Jaccard Index) between generated proposals and ground-truth annotations. For two boxes/regions b_i and b_j, IOU is defined as:

$$\text{IOU}(b_i, b_j) = \frac{area(b_i \cap b_j)}{area(b_i \cup b_j)} \tag{A.1}$$

   The following metrics are commonly used:

   • **Recall @ IOU Threshold** t: For each ground-truth instance, this metric checks whether the 'best' proposal from list L has IOU greater than a threshold t. If so, this ground truth instance is considered 'detected' or 'recalled'. Then average recall is measured over all the ground truth instances:

$$\text{Recall} \,@\, t = \frac{1}{|G|} \sum_{g_i \in G} \text{I}\,[\max_{l_j \in L} \text{IOU}(g_i, l_j) > t], \tag{A.2}$$

   where $I[\cdot]$ is an indicator function for the logical preposition in the argument. Object proposals are evaluated using this metric in two ways:

  - plotting Recall-*vs*.-#proposals by fixing t
  - plotting Recall-*vs*.-t by fixing the #proposals in L.

• **Area Under the recall Curve (AUC)**: AUC summarizes the area under the Recall-*vs*.-#proposals plot for different values of t in a single plot. This metric measures AUC-*vs*.-#proposals. It is also plotted by varying #proposals in L and plotting AUC-vs-t.

• **Volume Under Surface (VUS)**: This measures the average recall by linearly varying t and varying the #proposals in L on either linear or log scale. Thus it merges both kinds of AUC plots into one.

• **Average Best Overlap (ABO)**: This metric eliminates the need for a threshold. We first calculate the overlap between each ground truth annotation $g_i \in G$, and the 'best' object hypotheses in L. ABO is calculated as the average:

$$\text{ABO} = \frac{1}{|G|} \sum_{g_i \in G} \max_{l_j \in L} \text{IOU}(g_i, l_j) \tag{A.3}$$

ABO is typically is calculated on a per class basis. Mean Average Best Overlap (MABO) is defined as the mean ABO over all classes.

• **Average Recall (AR)**: This metric was recently introduced in [37]. Here, average recall (for IOU between 0.5 to 1)-*vs*.-#proposals in L is plotted. AR also summarizes proposal performance across different values of t. AR was shown to correlate with ultimate detection performance better than other metrics.

2. **Dataset**: The most commonly used datasets are the the PASCAL VOC [24] detection datasets. Note that these are *partially annotated* datasets where only the 20 PASCAL category instances are annotated. Recently analyses have been shown on ImageNet [38], which has more categories annotated than PASCAL, but is still a partially annotated dataset.

A.4 A Thought Experiment: How to Game the Evaluation Protocol

Let us conduct a thought experiment to demonstrate that the object proposal evaluation protocol can be 'gamed'.

Imagine yourself reviewing a paper claiming to introduce a new object proposal method – called DMP.

Before we divulge the details of DMP, consider the performance of DMP shown in Fig. A.3 on the PASCAL VOC 2010 dataset, under the AUC-*vs*.-#proposals metric.

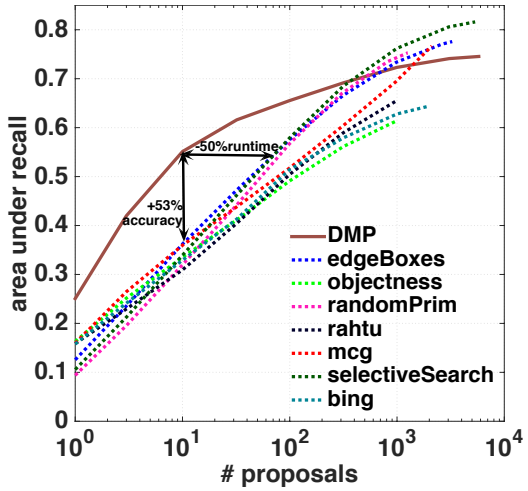


Figure A.3: Performance of different object proposal methods (dashed lines) and our proposed 'fraudulent' method (DMP) on the PASCAL VOC 2010 dataset. We can see that DMP *significantly* outperforms all other proposal generators. See text for details.

As we can clearly see, the proposed method DMP *significantly* exceeds all existing proposal methods [104, 22, 7, 2, 75, 63, 91, 14, 46] (which seem to have little variation over one another). The improvement at some points in the curve (*e.g.*, at M=10) seems to be *an order of magnitude* larger than all previous incremental improvements reported in the literature! In addition to the gain in AUC at a fixed M, DMPs also achieves the same AUC (0.55) at an *order of magnitude fewer* number of proposals (M=10 *vs.* M= 50 for edgeBoxes[104]). Thus, fewer proposals need to be processed by the ensuing detection system, resulting in an equivalent run-time speedup. This seems to indicate that a significant progress has been made in the field of generating object proposals.

So what is our proposed state-of-art technique DMP?
It is a mixture-of-experts model, consisting of 20 experts, where each expert is a deep feature (fc7)-based [20] objectness detector. At this point, you, the savvy reader, are probably already beginning to guess what we did.

DMP stands for 'Detector Masquerading as Proposal generator'. We trained object detectors for the 20 PASCAL categories (in this case with RCNN[31]), and then used these 20 detectors to produce the top-M most confident detections (after NMS), and declared them to be 'object proposals'.

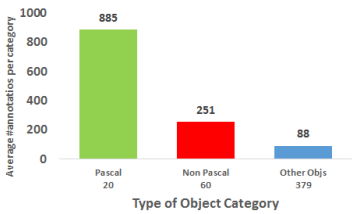
The point of this experiment is to demonstrate the following fact – clearly, no one would consider a collection of 20 object detectors to be a category independent object proposal method. However, our existing evaluation protocol declared the union of these top-M detections to be state-of-the-art.

Why did this happen? Because the protocol today involves evaluating a proposal generator

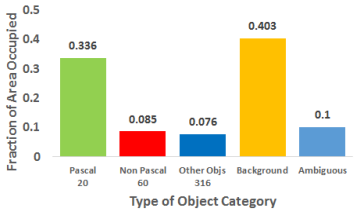
on *a partially annotated* dataset such as PASCAL. The protocol does not reward recall of non-PASCAL categories; in fact, early recall (near the top of the list of candidates) of non-PASCAL objects results in a penalty for the proposal generator! As a result, a proposal generator that tunes itself to these 20 PASCAL categories (either explicitly via training or implicitly via design choices or hyper-parameters) will be declared a better proposal generator when it may not be (as illustrated by DMP). Notice that as learning-based object proposal methods improve on this metric, "in the limit" *the best object proposal technique is a detector for the annotated categories*, similar to our DMP. Thus, we should be cautious of methods proposing incremental improvements on this protocol – improvements on this protocol do not necessarily lead to a better category independent object proposal method.

This thought experiment exposes the inability of the existing protocol to evaluate category independence.

A.5 Evaluation on Fully and Densely Annotated Datasets

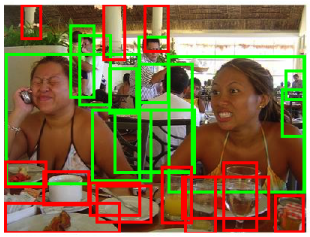


(a) Average #annotations for different categories.



(b) Fraction of image-area covered by different categories.

(c) PASCAL Context annotations [67].



(d) Our augmented annotations.

Figure A.4: (a),(b) Distribution of object classes in PASCAL Context with respect to different attributes. (c),(d) Augmenting PASCAL Context with instance-level annotations. (Green = PASCAL 20 categories; Red = new objects)

As described in the previous section, the problem of 'gameability' is occuring due to the

evaluation of proposal methods on partially annotated datasets. An intuitive solution would be evaluating on a *fully* annotated dataset.
In the next two subsections, we evaluate the performance of 7 popular object proposal methods[2, 91, 14, 104, 7, 75, 63] and two DMPs (RCNN[31] and DPM[26]) on one nearly-fully and two densely annotated datasets containing many more object categories. This is to quantify how much the performance of our 'fraudulent' proposal generators (DMPs) drops once the bias towards the 20 PASCAL categories is diminished (or completely removed).
We begin by *creating* a nearly-fully annotated dataset by building on the effort of PASCAL Context [67] and evaluate on this nearly-fully annotated modified instance level PASCAL Context; followed by cross-dataset evaluation on other partial-but-densely annotated datasets MS COCO [55] and NYU-Depth V2 [68].

Experimental Setup: On MS COCO and PASCAL Context datasets we conducted experiments as follows:

- Use the existing evaluation protocol for evaluation, *i.e.*, evaluate only on the 20 PASCAL categories.
- Evaluate on all the annotated classes.
- For the sake of completeness, we also report results on all the classes except the PASCAL 20 classes.[2]

Training of DMPs: The two DMPs we use are based on two popular object detectors - DPM[26] and RCNN[31]. We train DPM on 20 PASCAL categories and use it as an object proposal method. To generate large number of proposals, we chose a low value of threshold in Non-Maximum Suppression (NMS). Proposals are generated for each category and a score is assigned to them by the corresponding DPM for that category. These proposals are then merge-sorted on the basis of this score. Top M proposals are selected from this sorted list where M is the number of proposals to be generated.
Another (stronger) DMP is RCNN which is a detection pipeline that uses 20 SVMs (each for one PASCAL category) trained on deep features (fc7) [20] extracted on selective search boxes. Since RCNN itself uses selective search proposals, it should be viewed as a trained *reranker* of selective search boxes. As a consequence, it ultimately equals selective search performance once the number of candidates become large. We used the pretrained SVM models released with the RCNN code, which were trained on the 20 classes of PASCAL VOC 2007 trainval set. For every test image, we generate the Selective Search proposals using the 'FAST' mode and calculate the 20 SVM scores for each proposal. The 'objectness' score of a proposal is then the maximum of the 20 SVM scores. All the proposals are then sorted by this score and top M proposals are selected.[3]

[2]On NYU-Depth V2 performance is only evaluated on all categories. This is because only 8 PASCAL categories are present in this dataset.

[3]It was observed that merge-sorting calibrated/rescaled SVM scores led to inferior performance as compared to merge-sorting without rescaling.

Object Proposals Library: To ease the process of carrying out the experiments, we created an open source, easy-to-use object proposals library. This can be used to seamlessly generate object proposals using all the existing algorithms[104, 22, 7, 2, 75, 63, 76, 91, 40] (for which the Matlab code has been released by the respective authors and evaluate these proposals on any dataset using the commonly used metrics. This library will be made publicly available.

A.5.1 Fully Annotated Dataset

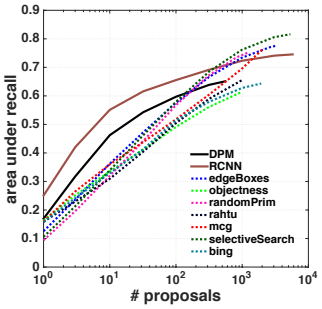
PASCAL Context: This dataset was introduced by Mottaghi *et al*. [67]. It contains additional annotations for all images of PASCAL VOC 2010 dataset [25]. The annotations are semantic segmentation maps, where *every single pixel* previously annotated 'background' in PASCAL was assigned a category label. In total, annotations have been provided for 459 categories. This includes the original 20 PASCAL categories and new classes such as keyboard, fridge, picture, cabinet, plate, clock.
Unfortunately, the dataset contains only category-level semantic segmentations. For our task, we needed instance-level bounding box annotations, which cannot be reliably extracted from category-level segmentation masks.

Creating Instance-Level Annotations for PASCAL Context: Thus, we created instance-level bounding box annotations for all images in PASCAL Context dataset. First, out of the 459 category labels in PASCAL Context, we identified 396 categories to be 'things', and ignored the remaining 'stuff' or 'ambiguous' categories[4] – neither of these lend themselves to bounding-box-based object detection. See supplement for details.
We selected the 60 most frequent non-PASCAL categories from this list of 'things' and manually annotated all their instances. Selecting only top 60 categories is a reasonable choice because the average per category frequency in the dataset for all the other categories (even after including background/ambiguous categories) was roughly one third as that of the chosen 60 categories (Fig. A.4a). Moreover, the percentage of pixels in an image left unannotated (as 'background') drops from 58% in original PASCAL to 50% in our nearly-fully annotated PASCAL Context. This manual annotation was performed with the aid of the semantic segmentation maps present in the PASCAL Context annotations. Examples annotations are shown in Fig. A.4d. For detailed statistics, see supplement.

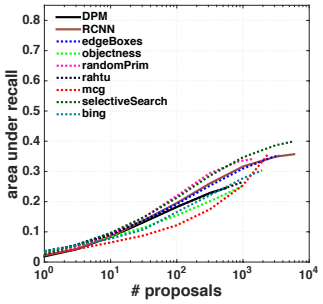
Results and Observations: We now explore how changes in the dataset and annotated categories affect the results of the thought experiment from Section A.4. Figs. A.5a, A.5b, A.5c, A.5h compare the performance of DMPs with a number of existing proposal methods [104, 22, 7, 2, 75, 63, 91, 14, 46] on PASCAL Context.

We can see in Column (a) that when evaluated on only 20 PASCAL categories DMPs trained on these categories appear to significantly outperform all proposal generators. However, we

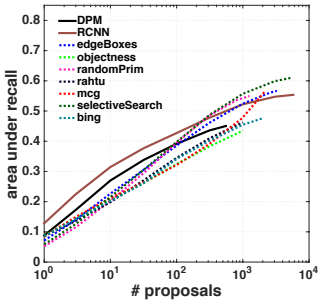
[4] *e.g.*, a 'tree' may be a 'thing' or 'stuff' subject to camera viewpoint.



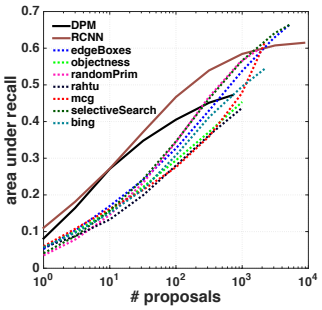
(a) Performance on PASCAL Context, only 20 PASCAL classes annotated.



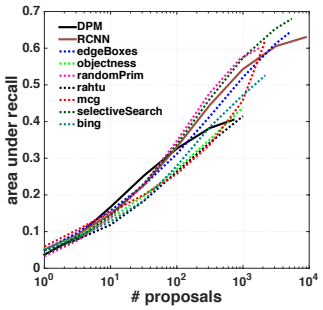
(b) Performance on PASCAL Context, only 60 non-PASCAL classes annotated.



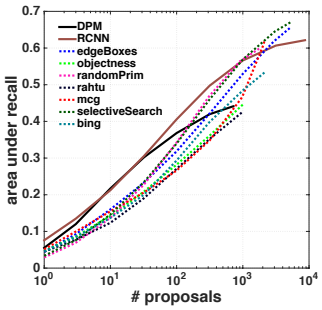
(c) Performance on PASCAL Context, all classes annotated.



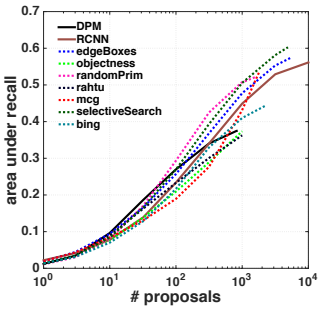
(d) Performance on MS COCO, only 20 PASCAL classes annotated.



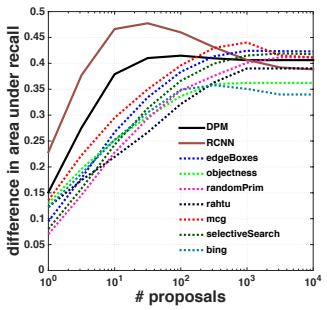
(e) Performance on MS COCO, only 60 non-PASCAL classes annotated.



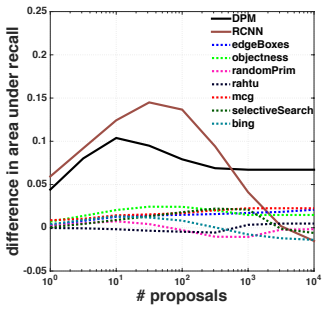
(f) Performance on MS COCO, all classes annotated.



(g) Performance on NYU-Depth V2, all classes annotated



(h) AUC @ 20 categories - AUC @ 60 categories on PASCAL Context.



(i) AUC @ 20 categories - AUC @ 60 categories on MS COCO.

Figure A.5: Performance of different methods on PASCAL Context, MS COCO and NYu Depth-V2 with different sets of annotations.

can see that they are not category independent because they suffer a big drop in performance when evaluated on 60 non-PASCAL categories in Column (b). Notice that on PASCAL context, *all proposal generators* suffer a drop in performance between the 20 PASCAL categories and 60 non-PASCAL categories. We hypothesize that this due to the fact that the non-PASCAL categories tend to be generally smaller than the PASCAL categories (which were the main targets of the dataset curators) and hence difficult to detect. But this could also be due to the reason that authors of these methods made certain choices while designing these approaches which catered better to the 20 annotated categories. However, the key observation here (as shown in Fig. A.5h) is that DMPs suffer the biggest drop. This drop is much greater than all the other approaches. It is interesting to note that due to the ratio of instances of 20 PASCAL categories vs other 60 categories, DMPs continue to slightly outperform proposal generators when evaluated on all categories, as shown in Column (c).

A.5.2 Densely Annotated Datasets

Besides being expensive, "full" annotation of images is somewhat ill-defined due to the hierarchical nature of object semantics (*e.g.* are object-parts such as bicycle-wheel, windows in a building, eyes in a face, *etc.* also objects?). One way to side-step this issue is to use datasets with dense annotations (albeit at the same granularity) and conduct cross-dataset evaluation.

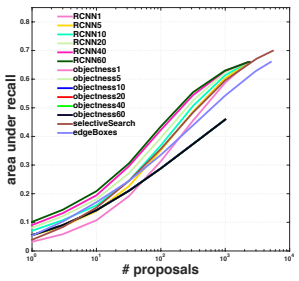
MS COCO: Microsoft Common Objects in Context (MS COCO) dataset [55] contains 91 common object categories with 82 of them having more than 5,000 labeled instances. It not only has significantly higher number of instances per category than the PASCAL, but also considerably more object instances per image (7.7) as compared to ImageNet (3.0) and PASCAL (2.3).

NYU-Depth V2: NYU-Depth V2 dataset [68] is comprised of video sequences from a variety of indoor scenes as recorded by both the RGB and Depth cameras. It features 1449 densely labeled pairs of aligned RGB and depth imageswith instance-level annotations. We used these 1449 densely annotated RGB images for evaluating object proposal algorithms. To the best of our knowledge, this is the first paper to compare proposal methods on such a dataset.

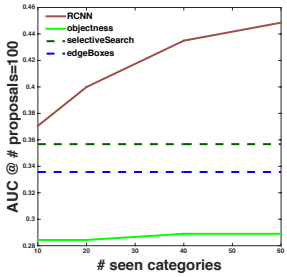
Results and Observations: Figs. A.5d, A.5e, A.5f, A.5i show a plot similar to PASCAL Context on MS COCO. Again, DMPs outperform all other methods on PASCAL categories but fail to do so for the Non-PASCAL categories. Fig. A.5g shows results for NYU-Depth V2. See that when many classes in the test dataset are not PASCAL classes, DMPs tend to perform poorly, although it is interesting that the performance is still not as poor as the worst proposal generators. Results on other evaluation criteria are in the supplement.

A.6 Bias Inspection

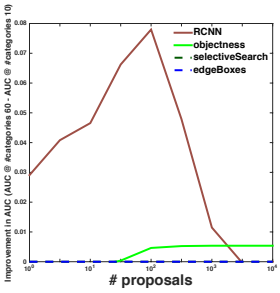
So far, we have discussed two ways of detecting 'gameability' – evaluation on nearly-fully annotated dataset and cross-dataset evaluations on densely annotated datasets. Although



(a) Area under recall *vs*. #proposals for various #seen classes



(b) Area under recall *vs*. #training-classes for #proposals = 100



(c) Improvement in area under recall from #seen classes =10 to 60 *vs*. #proposals.

Figure A.6: Performance of RCNN and other proposal generators vs number of object categories used for training. We can see that RCNN has the most 'bias capacity' while the performance of other methods is nearly (or absolutely) constant.

these methods are fairly useful for bias detection, they have certain limitations. Datasets can be unbalanced. Some categories can be more frequent than others while others can be hard to detect (due to choices made in dataset collection). These issues need to be resolved for perfectly unbiased evaluation. However, generating unbiased datasets is an expensive and time-consuming process. Hence, to detect the bias without getting unbiased datasets, we need a method which can measure performance of proposal methods in a way that category specific biases can be accounted for and the extent or the *capacity* of this bias can be measured. We introduce such a method in this section.

A.6.1 Assessing Bias Capacity

Many proposal methods[14, 46, 40, 47, 43, 50, 30, 72] rely on explicit training to learn an "objectness" model, similar to DMPs. Depending upon which, how many categories they are trained on, these methods could have a biased view of "objectness".
One way of measuring the *bias capacity* in a proposal method to plot the performance *vs*. the number of 'seen' categories while evaluating on some held-out set. A method that involves little or no training will be a flat curve on this plot. Biased methods such as DMPs will get better and better as more categories are seen in training. Thus, this analysis can help us find biased or 'gamebility-prone' methods like DMPs that are/can be tuned to specific classes. To the best of our knowledge, no previous work has attempted to measure bias capacity by varying the number of 'object' categories seen at training time. In this experiment,

we compared the performance of one DMP method (RCNN), one learning-based proposal method (Objectness), and two non learning-based proposal methods (Selective Search[91], EdgeBoxes[104]) as a function of the number of 'seen' categories (the categories trained on[5]) on MS COCO[55] dataset. Method names 'RCNNTrainN', 'objectnessTrainN' indicate that they were trained on images that contain annotations for only N categories (50 instances per category). Total number of images for all 60 categories was ˜2400 (because some images contain ¿1 object). Once trained, these methods were evaluated on a randomly-chosen set of ˜500 images, which had annnotations for all 60 categories.
Fig. A.6a shows Area under Recall *vs*. #proposals curve for learning-based methods trained on different sets of categories. Fig. A.6b and Fig. A.6c show the variation of AUC *vs*. # seen categories and improvement due to increase in training categories (from 10 to 60) *vs*. #proposals respectively, for RCNN and objectness when trained on different sets of categories. The key observation to make here is that with even a modest increase in 'seen' categories with the same amount of increased training data, performance improvement of RCNN is significantly more than objectness. Selective Search [91] and edgeBoxes [104] are the dashed straight lines since there is no training involved.

These results clearly indicate that as RCNN sees more categories, its performance improves. One might argue that the reason might be that the method is learning more 'objectness' as it is seeing more data. However, as discussed above, the increase in the dataset size is marginal (˜40 images per category) and hence it unlikely that such a significant improvement is observed due to that. Thus, it is reasonable to conclude that this improvement is because the method is learning class specific features.

Thus, this approach can be used to reason about 'gameability-prone' and 'gameability-immune' proposal methods without creating an expensive fully annotated dataset. We believe this simple but effective diagnostic experiment would help to detect and thus contribute in managing the category specific bias in all learning-based methods.

A.7 Conclusion

In this work, we make an explicit distinction between the two mutually co-existing but different interpretations of object proposals. The current evaluation protocol for object proposal methods is suitable only for detection proposals and is a biased 'gameable' protocol for category-independent object proposals. By evaluating only on a specific set of object categories, we fail to capture the performance of the proposal algorithm on all the remaining object categories that are present in the test set, but not annotated in the ground truth. We demonstrate this gameability via a simple thought experiment where we propose a 'fraudulent' object proposal method that outperforms all existing object proposal techniques on

[5]The seen categories are picked in the order they are listed in MS COCO dataset (*i.e.*, no specific criterion was used).

current metrics. We conduct a thorough evaluation of existing object proposal methods on three densely annotated datasets. We introduce a fully-annotated version of PASCAL VOC 2010 where we annotated all instances of all object categories occurring in all images. We hope this dataset will be broadly useful.

Furthermore, since densely annotating the dataset is a tedious and costly task; we proposed a set of diagnostic tools to plug the vulnerability of the current protocol.

Fortunately, we find that none of existing proposal methods seem to be biased, most of the existing algorithms and do generalize well to different datasets and in our experiments even on densely annotated datasets. In that sense, our findings are consistent with results in [37]. However, that should not prevent us from recognizing and safeguarding against the flaws in the protocol, lest we over-fit as a community to a specific set of object classes.

Bibliography

[1] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi. C-vqa: A compositional split of the visual question answering (vqa) v1.0 dataset. *CoRR*, 2017.

[2] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *PAMI*, 2012.

[3] P. Anderson, B. Fernando, M. Johnson, and S. Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer, 2016.

[4] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 39–48, 2016.

[5] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015.

[6] P. Arbelaez, B. Hariharan, C. Gu, S. Gupta, L. D. Bourdev, and J. Malik. Semantic segmentation using regions and parts. 2012.

[7] P. Arbeláez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. 2014.

[8] D. Banica and C. Sminchisescu. CPMC-3D-O2P: semantic segmentation of RGB-D images using CPMC and second order pooling. *CoRR*, abs/1312.7715, 2013.

[9] D. Banica and C. Sminchisescu. Second-order constrained parametric proposals and sequential search-based structured prediction for semantic segmentation in rgb-d images. In *CVPR*, 2015.

[10] J. Bos and K. Markert. Recognising textual entailment with logical inference. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 628–635, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[11] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. 2012.

[12] J. Carreira and C. Sminchisescu. CPMC: Automatic Object Segmentation Using Constrained Parametric Min-Cuts. *PAMI*, 34, 2012.

[13] X. Chen, H. Ma, X. Wang, and Z. Zhao. Improving object proposals with multi-thresholding straddling expansion. In *CVPR*, 2015.

[14] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr. Bing: Binarized normed gradients

for objectness estimation at 300fps. In *CVPR*, 2014.
[15] M. Cho, S. Kwak, C. Schmid, and J. Ponce. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. *CoRR*, abs/1501.06170, 2015.
[16] R. G. Cinbis and S. Sclaroff. Contextual object detection using set-based classification. 2012.
[17] R. G. Cinbis, J. Verbeek, and C. Schmid. Segmentation Driven Object Detection with Fisher Vectors. 2013.
[18] J. Dai, K. He, and J. Sun. Convolutional feature masking for joint object and stuff segmentation. In *CVPR*, 2015.
[19] T. Deselaers, B. Alexe, and V. Ferrari. Weakly supervised localization and learning with generic knowledge. (3):275–293, 2012.
[20] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.
[21] M. Elhoseiny, S. Cohen, W. Chang, B. Price, and A. Elgammal. Automatic annotation of structured facts in images. *arXiv preprint arXiv:1604.00466*, 2016.
[22] I. Endres and D. Hoiem. Category-independent object proposals with diverse ranking. *PAMI*, 36(2):222–234, 2014.
[23] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov. Scalable object detection using deep neural networks. *CoRR*, abs/1312.2249, 2013.
[24] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.
[25] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html.
[26] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9):1627–1645, 2010.
[27] K. Fragkiadaki, P. Arbelaez, P. Felsen, and J. Malik. Spatio-temporal moving object proposals. *arXiv preprint arXiv:1412.6504*, 2014.
[28] 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. *arXiv preprint arXiv:1606.01847*, 2016.
[29] 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 answering. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS'15, pages 2296–2304, Cambridge, MA, USA, 2015. MIT Press.
[30] A. Ghodrati, A. Diba, M. Pedersoli, T. Tuytelaars, and L. Van Gool. Deepproposal: Hunting objects by cascading deep convolutional layers. 2015.
[31] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524, 2013.
[32] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik. Learning Rich Features from RGB-D

Images for Object Detection and Segmentation. *arXiv preprint arXiv:1407.5736*, 2014.
[33] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *CoRR*, 2014.
[34] S. He and R. W. Lau. Oriented object proposals. In *ICCV*, 2015.
[35] G. Heitz and D. Koller. Learning spatial context: Using stuff to find things. In *Proc. 10th European Conference on Computer Vision*, 2008.
[36] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, S. Kate, and T. Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
[37] J. Hosang, R. Benenson, P. Dollár, and B. Schiele. What makes for effective detection proposals? *arXiv preprint arXiv:1502.05082*, 2015.
[38] J. Hosang, R. Benenson, and B. Schiele. How good are detection proposals, really? In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014.
[39] J.-H. Huang, M. Alfadly, and B. Ghanem. VQABQ: Visual Question Answering by Basic Questions. *ArXiv e-prints*, 2017.
[40] A. Humayun, F. Li, and J. M. Rehg. Rigor- recycling inference in graph cuts for generating object regions. In *CVPR*, 2014.
[41] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei. Image retrieval using scene graphs. In *Conference on Computer Vision and Pattern Recognition*, pages 3668–3678, 2015.
[42] C. Kading, A. Freytag, E. Rodner, P. Bodesheim, and J. Denzler. Active learning and discovery of object categories in the presence of unnameable instances. In *CVPR*, 2015.
[43] H. Kang, M. Hebert, A. A. Efros, and T. Kanade. Data-driven objectness. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(1):189–195, 2015.
[44] A. Karpathy and F.-F. Li. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.
[45] J.-H. Kim, S.-W. Lee, D. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang. Multimodal residual learning for visual qa. In *Advances in Neural Information Processing Systems*, pages 361–369, 2016.
[46] P. Krähenbühl and V. Koltun. Geodesic object proposals. 2014.
[47] P. Krähenbühl and V. Koltun. Learning to propose objects. In *CVPR*, 2015.
[48] P. Krähenbühl and V. Koltun. Learning to propose objects. In *CVPR*, 2015.
[49] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016.
[50] W. Kuo, B. Hariharan, and J. Malik. Deepbox: Learning objectness with convolutional networks. 2015.
[51] A. Kuznetsova, S. Ju Hwang, B. Rosenhahn, and L. Sigal. Expanding object detector's horizon: Incremental learning framework for object detection in videos. In *CVPR*, 2015.
[52] S. Kwak, M. Cho, I. Laptev, J. Ponce, and C. Schmid. Unsupervised object discovery

and tracking in video collections. *CoRR*, abs/1505.03825, 2015.

[53] A. Lai and J. Hockenmaier. Learning to predict denotational probabilities for modeling entailment. In *EACL*, 2017.

[54] M. Lin, Q. Chen, and S. Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

[55] 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.

[56] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014.

[57] 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.

[58] J. Lu, C. Xiong, D. Parikh, and R. Socher. Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning. *CVPR*, 2017.

[59] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297, 2016.

[60] L. Ma, Z. Lu, and H. Li. Learning to answer questions from image using convolutional neural network. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 3567–3573. AAAI Press, 2016.

[61] M. Malinowski and M. Fritz. 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, 2014.

[62] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–9, 2015.

[63] S. Manen, M. Guillaumin, and L. Van Gool. Prime object proposals with randomized prim's algorithm. 2013.

[64] M. Marelli, S. Menini, M. Baroni, L. Bentivogli, R. Bernardi, and R. Zamparelli. A sick cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), 2014.

[65] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[66] I. Misra, A. Shrivastava, and M. Hebert. Watch and learn: Semi-supervised learning for object detectors from video. In *CVPR*, 2015.

[67] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 2014.

[68] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[69] D. Oneata, J. Revaud, J. Verbeek, and C. Schmid. Spatio-temporal object detection

proposals. In *ECCV 2014*, 2014.

[70] W. Ouyang, P. Luo, X. Zeng, S. Qiu, Y. Tian, H. Li, S. Yang, Z. Wang, Y. Xiong, C. Qian, et al. Deepid-net: multi-stage and deformable deep convolutional neural networks for object detection. *arXiv preprint arXiv:1409.3505*, 2014.

[71] D. H. Park, L. A. Hendricks, Z. Akata, B. Schiele, T. Darrell, and M. Rohrbach. Attentive explanations: Justifying decisions and pointing to the evidence. *arXiv preprint arXiv:1612.04757*, 2016.

[72] P. O. Pinheiro, R. Collobert, and P. Dollár. Learning to segment object candidates. *CoRR*, abs/1506.06204, 2015.

[73] J. Pont-Tuset and L. Van Gool. Boosting object proposals: From pascal to coco. In *International Conference on Computer Vision*, 2015.

[74] A. Prest, C. Schmid, and V. Ferrari. Weakly supervised learning of interactions between humans and objects. *PAMI*, 34(3):601–614, 2012.

[75] E. Rahtu, J. Kannala, and M. B. Blaschko. Learning a category independent object detection cascade. 2011.

[76] P. Rantalankila, J. Kannala, and E. Rahtu. Generating object segmentation proposals using global and local search. 2014.

[77] A. Ray, G. Christie, M. Bansal, D. Batra, and D. Parikh. Question relevance in vqa: Identifying non-visual and false-premise questions. *arXiv preprint arXiv:1606.06622*, 2016.

[78] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015.

[79] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu. Unsupervised joint object discovery and segmentation in internet images. In *CVPR*. IEEE, 2013.

[80] O. Russakovsky, J. Deng, J. Krause, A. Berg, and L. Fei-Fei. The ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012). http://www.image-net.org/challenges/LSVRC/2012/, 2012.

[81] 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, 2014.

[82] B. Sacaleanu, C. Orasan, C. Spurk, S. Ou, O. Ferrandez, M. Kouylekov, and M. Negri. Entailment-based question answering for structured data. In *22Nd International Conference on on Computational Linguistics: Demonstration Papers*, COLING '08, pages 173–176, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[83] M. Sammons, V. G. V. Vydiswaran, and D. Roth. "ask not what textual entailment can do for you...". In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1199–1208, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[84] S. Schuster, R. Krishna, A. Chang, L. Fei-Fei, and C. D. Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the Fourth Workshop on Vision and Language*, pages 70–80, 2015.

[85] F. Sener, C. Bas, and N. Ikizler-Cinbis. On recognizing actions in still images via multiple features. 2012.

[86] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[87] J. Sun and H. Ling. Scale and object aware image retargeting for thumbnail browsing. 2011.

[88] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014.

[89] C. Szegedy, S. Reed, D. Erhan, and D. Anguelov. Scalable, high-quality object detection. *CoRR*, abs/1412.1441, 2014.

[90] Y.-H. Tsai, O. C. Hamsici, and M.-H. Yang. Adaptive region pooling for object detection. In *CVPR*, 2015.

[91] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. *IJCV*, 2013.

[92] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun. Order-embeddings of images and language. In *ICLR*, 2016.

[93] C. Wang, L. Zhao, S. Liang, L. Zhang, J. Jia, and Y. Wei. Object proposal by multi-branch hierarchical segmentation. In *CVPR*, 2015.

[94] N. Wang, S. Li, A. Gupta, and D. Yeung. Transferring rich feature hierarchies for robust visual tracking. *CoRR*, abs/1501.04587, 2015.

[95] P. Wang, Q. Wu, C. Shen, A. v. d. Hengel, and A. Dick. Fvqa: Fact-based visual question answering. *arXiv preprint arXiv:1606.05433*, 2016.

[96] X. Wang, M. Yang, S. Zhu, and Y. Lin. Regionlets for generic object detection. 2013.

[97] Q. Wu, C. Shen, L. Liu, A. Dick, and A. van den Hengel. What value do explicit high level concepts have in vision to language problems? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 203–212, 2016.

[98] Z. Wu, F. Li, R. Sukthankar, and J. M. Rehg. Robust video segment proposals with painless occlusion handling. In *CVPR*, 2015.

[99] D. Xu, Y. Zhu, C. Choy, and L. Fei-Fei. Scene graph generation by iterative message passing. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[100] G. Yu and J. Yuan. Fast action proposals for human action detection and search. In *CVPR*, 2015.

[101] Y. Zhang, X. Chen, J. Li, C. Wang, and C. Xia. Semantic object segmentation via detection in weakly labeled video. In *CVPR*, 2015.

[102] Y. Zhu, R. Urtasun, R. Salakhutdinov, and S. Fidler. segdeepm: Exploiting segmentation and context in deep neural networks for object detection. In *CVPR*, 2015.

[103] C. L. Zitnick, A. Agrawal, S. Antol, M. Mitchell, D. Batra, and D. Parikh. Measuring machine intelligence through visual question answering. *arXiv preprint arXiv:1608.08716*, 2016.

[104] C. L. Zitnick and P. Dollar. Edge boxes: Locating object proposals from edges. 2014.