Question Relevance in Visual Question Answering

Prakruthi Prabhakar
ML with Large Datasets (10-805)
Carnegie Mellon University
Pittsburgh, PA
prakrutp@andrew.cmu.edu

Nitish Kulkarni
ML with Large Datasets (10-805)
Carnegie Mellon University
Pittsburgh, PA
nitishkk@andrew.cmu.edu

Linghao Zhang
ML with Large Datasets (10-805)
Carnegie Mellon University
Pittsburgh, PA
linghaoz@andrew.cmu.edu

ABSTRACT

Free-form and open-ended Visual Question Answering systems solve the problem of providing an accurate natural language answer to a question pertaining to an image. In this paper, we solve the problem of identifying the relevance of the posed question to the image. We address the problem as two sub-problems. We first identify if the question is visual or not. If the question is visual, we then determine if it's relevant to the image or not. We present the results of two models to identify if the question is visual. We also present the data extraction methodology for solving the relevance to the image, given a visual question. We aim to solve the second sub-problem as part of the final project.

1 INTRODUCTION

The task of automatically answering questions in the context of visual information has gained prominence in the last few years. Being able to answer open-ended questions about an image is a challenging task, but one of great practical significance. For instance, visually impaired individuals might inquire about different aspects of an image in the form of free-form questions. However, when Visual Question Answering (VQA) systems are provided with irrelevant questions, they tend to provide nonsensical answers. VQA systems in real world scenarios are expected to be sophisticated to identify the relevance of posed free-form questions to the input image, to better answer them.

There are two aspects of relevance of a question to the input image:

(1) Non-visual questions which do not require any input image to answer the question
(2) False-premise question which require an input image but do not pertain to the provided input image

In this project, we formulate the problem as follows:

> Given an image and a natural language question about the image, identify if the question is relevant to the input image.

For visual versus non-visual question detection, we present the results of two approaches. The first approach is based on training a Logistic Regression model using unigrams, bigrams and trigrams of Part-of-Speech (POS) tags of the question. In the second approach, we use a Long Short-Term Memory (LSTM) Recurrent Neural Network trained on Part-of-Speech (POS) tags to capture linguistic structure of questions.

For the second sub-problem of identifying true versus false premise of a visual question to an image, we use the data extraction methodology in [9] to obtain question-image pairs for true and false premise. We also present the baselines used as reference for both the problems.

2 RELATED WORK

There have been significant advances in recent years on identifying the similarity between images and textual information. Text-based image retrieval [8] systems and visual semantic alignments in image captioning models [7], [4] are some examples of efforts in that direction. While some systems do not answer when the input is ill-formed or likely to result in failure, some others try to find the most meaningful answer to such inputs. [3] tries separating visual text from non-visual text from descriptions of images and use it for enhancing image captioning systems. These ideas can be used to boost the performance of visual question relevance task.

Much of our work is based on the problem and approaches presented in [10]. In this paper, the authors identify the two facets of question relevance for Visual Question Answering, i.e. categorizing the questions as visual versus non-visual questions, and then identifying if a question has a true premise for an image. The paper also provides several baselines for both the problems. For the visual versus non-visual classification, the authors propose a heuristic-based and an LSTM-based approach. And for the true versus false premise problem, there are three baselines based on entropy from VQA models, question-caption similarity and question-question similarity.

For the problem to identifying false premise, [9] also makes significant contribution for extracting the premise from a question. The authors of this work also go on to create a well-curated, much larger and more class-balanced dataset for the true versus false premise based on the VQA dataset [2]. In this paper, the concept of a premise in a question is explored in greater detail and a more diverse set of problems are addressed, such as - given an image and a question with a false premise, can we predict the premise in the question that cannot be answered by the image. However, the focus of our work is on exploring scalable algorithms and architectures for answering *if* the question can be answered in the context of the image.

In the context of establishing semantic relationships between images and text, the work done in [1] is also quite relevant, where the authors of this paper propose a holistic embedding technique for combining multiple and diverse language representations for better transfer of knowledge, by mapping visual and language parts into a common embedding space. Although our problem is supervised, we believe that such an embedding could help improve the identification of question relevance for an image.

3 DATASET

For the first task of detecting visual versus non-visual questions, we refer to the methodology used in [10]. Since VQA 2.0 dataset [5] is now available, we use the training, validation and test questions for images from this dataset as visual questions. For non-visual questions, we use the philosophical and general knowledge questions provided by [10]. Combining the two sources and eliminating duplicate questions, we have 160,010 questions for training, 82,067 questions for validation and 148,927 questions for test datasets.

The dataset for the second sub-problem of detecting questions with false premise is based on the VQA corpus [2]. The data acquisition involves creating image-question pairs with true and false premises for the question. The questions in the VQA dataset can be assumed to have true premises, since they were manually generated.

To identify the questions with false premises, we choose to use the questions from the same dataset, but for other images. Here, we explore three approaches:

(1) **Question Similarity**
In this approach, for every image, we use the set of true-premise questions from the VQA dataset [2], and extract k least similar questions from the set. As similarity measure, we try using doc2vec similarity and word2vec similarity for keywords in a question (nouns, verbs and adjectives).

(2) **Visual True vs False Premise Questions (VTFQ) Dataset**
In this approach, we use the dataset presented in [10], where the methodology is similar to the one described for question similarity, but instead of using a question similarity measure, the authors sample random questions from the set, and have AMT workers annotate the questions as relevant and not-relevant to the corresponding images. The VTFQ dataset consists of 10,793 question-image pairs with 1,500 unique images of which 79% of the pairs have false premise.

(3) **Question Relevance Prediction and Explanation (QRPE) Dataset**
The problem with the former two approaches is that using random or least similar questions for an image would make the problem of false-premise detection much easier than the case where the a single premise among all the premises of the question were to be false in the context of the image. To generate such a dataset, we employ the approach described in [9]. This approach first outlines a premise-extraction methodology for a given question. Then, for a given question-image pair in the VQA dataset, a set of all images is created which has exactly one premise as false for the question, which are referred as negative images. To make the false-premise detection problem challenging, we pick the negative images that are most similar to the positive image (i.e. image with true premise for the question). In [2], the QRPE dataset consists of 53,911 tuples of the form (I^+, Q, P, I^-), where (I^+, Q) is a pair of positive image and true premise question, I^- is the negative image with P as the premise in Q that is false for I^-. While a single QI pair for true and false premises is extracted using this approach, we extend this methodology to create a larger dataset by using more QI pairs with true premise, and having multiple negative images for a single question.

Table 1: Results of visual versus non-visual question detection using different POS features in Logistic Regression. Here, Uni represents using unigram POS tags as features, Uni+Bi represents using unigram as well as bigram POS tags as features, Uni+Bi+Tri represents using unigram, bigram and trigram as features.

Metric	Uni	Uni+Bi	Uni+Bi+Tri
Precision (Visual class)	0.9990	0.9996	0.9997
Recall (Visual class)	0.9980	0.9997	0.9993
Precision (Non-visual class)	0.8193	0.9690	0.9354
Recall (Non-visual class)	0.9049	0.9638	0.9705

Table 2: Comparison of results of Logistic Regression and LSTM based models for visual versus non-visual question detection

Metric	Logistic Regression	LSTM
Precision (Visual class)	0.9996	0.9995
Recall (Visual class)	0.9997	1.0
Precision (Non-visual class)	0.9690	1.0
Recall (Non-visual class)	0.9638	0.9511

4 APPROACH

We present the approaches used for visual-vs-non-visual question detection.

4.1 Visual vs. Non-visual Question Detection

We identify that non-visual questions have different linguistic structure than visual questions. For example, non-visual questions such as "Name the national rugby team of Argentina." or "Who is the president of Zimbabwe?" often have differences in structure in comparison to visual questions such as "Is this truck yellow?" and "What color are the giraffes?". Hence, we use Spacy [6] to process all questions to obtain Part-of-Speech (POS) tags as features. We compare two models, a Logistic Regression model versus an LSTM-RNN based approach.

(1) **Logistic Regression** We trained a Logistic Regression model using POS tags of the questions as features. We also experimented with larger feature sets using bi-grams and tri-grams of POS tags. We implemented a streaming scalable version of logistic regression for training. We assume that the validation and test datasets fit in memory for this problem.

(2) **LSTM** We also trained an LSTM model using the architecture from [10] for modeling visual vs. non-visual question detection. This model uses a dimensionality of 100 for hidden vectors and POS tags of the words in the question as the input. We also experimented with alternate architectures with varying dimensionality of the hidden vectors as well as POS-tag embeddings.

5 EXPERIMENTS AND RESULTS

The results for visual vs. non-visual models are presented in Table 1 and Table 2. Since there is a class imbalance problem in the datasets, we report the average per-class (i.e., normalized) metrics for all approaches.

Table 1 compares the results for the three models of Logistic Regression, using unigrams, bigrams and trigrams of POS tags as features. As can be observed from the table, addition of bigrams of POS tags as features helped improve the precision and recall of both classes significantly in comparison to using only unigrams of POS tags as features. However, using trigrams as additional features didn't give significant improvement in the metrics. Additionally, using trigrams as features increases computational time exponentially. Hence, we use unigram and bigram based logistic regression to compare with results from LSTM model.

We trained several models of LSTM using varying embedding and hidden vector dimensionality using NVIDIA Tesla K80 GPU. We observed that all models performed similarly across different metrics. Hence, we provide the results of replicating the model provided in [10]. Since we use VQA 2.0 dataset [5] and different set of POS tags from spacy, we identify that the results are different from the original paper for the same model.

From table 2, we observe that logistic regression performs better than LSTM based approach for some metrics and performs comparatively for others. However, we provide a streaming scalable implementation of logistic regression, which takes significantly lesser time for training in comparison to LSTM.

6 CONCLUSION

In this project, we attempt the problem of identifying relevance of posed questions to visual question answering systems by exploring more scalable approaches that yield similar or better results (by virtue of larger training data). For the first sub-problem of identifying visual versus non-visual question detection, we provide a time-efficient and scalable implementation of logistic regression. This approach provides comparative or better results on all metrics in comparison to strong baselines provided by LSTM based approach. We also have implemented a data extraction pipeline for obtaining larger data for the second sub-problem of true versus false premise detection. We aim to solve the second sub-problem as part of the final report using LSTM-based architectures as well as question-image embeddings.

REFERENCES

[1] Zeynep Akata, Mateusz Malinowski, Mario Fritz, and Bernt Schiele. 2016. Multi-Cue Zero-Shot Learning with Strong Supervision. *CoRR* abs/1603.08754 (2016). arXiv:1603.08754 http://arxiv.org/abs/1603.08754

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. *CoRR* abs/1505.00468 (2015). arXiv:1505.00468 http://arxiv.org/abs/1505.00468

[3] Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Karl Stratos, Kota Yamaguchi, Yejin Choi, Hal Daumé III, Alexander C Berg, et al. 2012. Detecting visual text. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 762–772.

[4] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1473–1482.

[5] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. *arXiv preprint arXiv:1612.00837* (2016).

[6] Matthew Honnibal and Mark Johnson. 2015. An Improved Non-monotonic Transition System for Dependency Parsing. In *EMNLP*.

[7] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3128–3137.

[8] Dong Liu, Xian-Sheng Hua, Meng Wang, and HongJiang Zhang. 2009. Boost search relevance for tag-based social image retrieval. In *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*. IEEE, 1636–1639.

[9] Aroma Mahendru, Viraj Prabhu, Akrit Mohapatra, Dhruv Batra, and Stefan Lee. 2017. The promise of premise: Harnessing question premises in visual question answering. *arXiv preprint arXiv:1705.00601* (2017).

[10] Arijit Ray, Gordon Christie, Mohit Bansal, Dhruv Batra, and Devi Parikh. 2016. Question relevance in VQA: identifying non-visual and false-premise questions. *arXiv preprint arXiv:1606.06622* (2016).