

Pragmatic Visual Question Answering on Color Images

Mathew Mathew

mathewma@gmail.com

Abstract

Presented in this work is a new Visual Question Answering(VQA) Dataset and an associated model, that answers questions around Color Images. Correctly answering the question requires the model to show understanding of the underlying context ie Pragmatics or Grounding. This dataset and model is targeted at Machine Learning Students/Enthusiasts trying to build their intuitions around DeepLearning, VQA concepts, Multi-Task learning and MultiModal Architectures. The ColorImages dataset is a simple but still challenging dataset because the language of color is rich and nuanced. However it is easier to reason compared to the complexity around models built around VQA datasets from (Agrawal et al., 2015) and (Hudson and Manning, 2019) which are more research oriented.

1 Introduction

The ability to understand a visual environment and reason and communicate around it, is a very life like quality. This is an ability that humans have developed and extended even further by being able to describe their environment in very rich language.

Visual Question Answering (VQA) represents an AI task that perhaps most closely emulates this ability. The models addressing this task will have to build on work from different domain areas such as Computer Vision (CV), NLP (Natural Language Processing) and Knowledge Representation Reasoning (KR) and tie them together. The models will also have show common sense understanding, ie grounding to be able to correctly answers the questions.

It involves reasoning around image content and answering open ended questions such as "how many trucks are on the road". Solving this problem requires building a generalized common model,

that can solve a number of different problems that include: Identifying objects in a scene and relationship between them, representing it language, linking the visual and language modality so that reasoning can be applied and last but not the least having common knowledge (pragmatics) that extends beyond the information that has been presented.

The AI tasks, we look to solve is to build a model that can communicate a selected color among a group of colors embedded in a color image. Where the selected color in the image is marked by a bounding box. It is important to emphasize here that this marking is something that humans can automatically intuit because of the common contextual knowledge we have. We look to see if we can build this intuition in the model.

This work extends the work of (Monroe et al., 2017) where they present in their paper "Colors in Context: A Pragmatic Neural Model for Grounded Language Understanding" a Pragmatic Language Model that can successfully communicate a "selected color" from a group of colors between a speaker and a listener, by re-framing it as a Visual Question Answering Task.

We hypothesis that their AI task can be successfully tackled using of a generic MultiModal Architecture and within the framework of a VQA task. Our Image VQA dataset (Table-1) is derived from their dataset. Their work also serves us as a baseline because we use the same Learning Accuracy metric and evaluation dataset to evaluate our Model.

2 Related Work

Solving VQA tasks requires building a generalized common model, that can solve a number of different problems such as identifying objects and the

Question Type	Example
Object detection	how many bikes are there
Fine-grained recognition	what kind of cheese is on the pizza
Activity recognition	is the man crying

Table 1: Examples of the type of questions that are intended to be answered by the (Agrawal et al., 2015) dataset

relationship between the objects in a scene, representing the scene in language, linking the visual and language modality so that reasoning can be applied and last but not the least having common knowledge (pragmatics) that extends beyond the information that has been presented.

This type of problem is a natural fit for Multi-Modal neural models as can be seen in some of the referenced papers below.

MultiModal learning introduced by (Kaiser et al., 2017) showed how a single model can be used to learn representations from different domain areas while still versatile enough to solve a number of different AI tasks. The tasks solved by the presented model include image classification, image captioning, language translation and speech recognition. It trains against 8 different corpora that include English text, ImageNet Day, COCO image captioning data, parsing data, English-German translation and English-French translation to execute on the tasks listed above. The benchmark results are comparable to specialized models created for a particular functional domain. The authors say that an advantage of this approach is that it can be used to compensate for areas where there is not enough data as the model can still fall back on what it learned from the other sources.

Question Answering (VQA) introduced by (Agrawal et al., 2015) by presented a free-form and open ended Visual Questioning Answering task (VQA). The aim of the task is to answer open-end questions on a provided image in accurate natural language. Through their work they established a quantitative benchmark against with VQA models can be evaluated.

GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering In this paper (Hudson and Manning, 2019), a newer vqa dataset called *gqa* is introduced. This dataset based on the *Visual Genome dataset* looks to address issues in prior data vqa data sets. In particular it aims at reducing the bias inherent

in prior data sets, which allows the model to cheat, that is answer questions without looking at the image data. For example a model might be able to guess that an apple is red or green without actually looking at the image. The Visual Genome data set is an annotated image data-set in which objects in the image are identified and the relationships between them are captured in a knowledge graph. Using computation linguistic methods, question and answer sets are then generated from the knowledge graph captured in the Visual Genome data-set. It results in a more precise or less open ended question answer set, that requires understanding of the compositional interdependence of objects in the image to answer properly. The paper also introduces the following new metrics in evaluating vqa models - Consistency, Validity, Plausibility and Grounding.

Cross-Modality Encoder Representations from Transformers(LXMERT) introduced by (Tan and Bansal, 2019) is a multimodal neural model that benchmarks well against the VQA dataset. As is the norm with current state of the art models, it makes use of transformers, however the underlying model architecture is similar in concept to what was presented in (Kaiser et al., 2017). In this model the inputs the (Questions and Images) are encoded separately using different neural architectures Convnets for Images, and transformers for the Questions and then combined in what can be considered as the Encoder part of model based on the EncoderDecoder Architecture.

Colors in Context: A Pragmatic Neural Model for Grounded Language Understanding by (Monroe et al., 2017) introduces a model and a dataset that ties together color and the language of color. The Model looks to to effectively describe a selected color from a group of other colors to enable a listener infer the selection based on what was described. The generated language should be rich enough for a listener to be able to discern between nuanced shades of a particular color. Examples of such descriptions include: light grey blue, pinkish


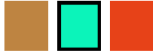


Image	Question	Answer
	Describe the selected color	Deep Pink
	Describe the selected color	Blueish Teal
	Describe the selected color	Blueish Teal
	Describe the color selected by the black border	Blueish Grey

Table 2: Sample of data from the introduced VQA Image Color dataset. Is made up of images, some of them are selected by a border. The model has to answer questions, related to the color that was selected. Note the nuance in the 4th image: two of the colors are surrounded by borders and that the question to be answered is to selected the color bounded by the black border.

purple.

Doing this successfully requires the model to factor in pragmatics so that the color is described in the context of the environment. The implemented model makes use of feature engineering techniques such as fourier transformation to create a color embedding. A decoder network is then is than trained against these embeddings and color sentence pairs, factoring through attention the actual color that was selected.

3 Data

The data for this experiment is synthesized from the [SCM colors dataset](#) presented in the (Monroe et al., 2017). Each record in that dataset is extended by generating image variations of the colors and an associated question as illustrated in Table-1. The raws-colors in the original dataset are used to generate image files that contain variations of the same data, with the selected color in different positions. Similarly questions are generated.

The created dataset is organized in a csvfile and an associated image folder. The fields in the csvfile are described below.

DataField	Description
id	Data-identifer
image-id	image-identifier
question	Question
condition	Measure of challenge
Answer	Answer

Condition is a measure of how challenging the color is to describe.

1. close, represents color images that shades of each other

2. split, where one distractor image is a shade of the target but the other distractor image is more further away.

3. far, represents images that are far apart.

We also introduce additional subtleties to make the the dataset more interesting, such as generating additional borders, as can be seen in the 4th example in Table-1, where one border is selected by a black border and another color is selected by a green border. This allows for asking more nuanced questions against the dataset, such as what color is selected by the green border.

This augmentation approach results in quadrupling the SCM dataset from 57947 rows to 231788 rows.

4 Model

The model as illustrated in Figure 1, is based on the EncoderDecoder approach used by MutliModal models referenced before. It creates a common embedding layers from the two inputs the Images and Questions which are encoded in their respective ImageEncoder and QuestionEncoder networks. This combined embedding is then feed into the hidden layer of the AnswerDecoder, which is a GRU based RNN.

The ImageEncoder is a simple Convnet network, that is composed of a couple of Convnet, Pooling, Linear and ReLU layers. The QuestionEncoder is another GRU based RNN

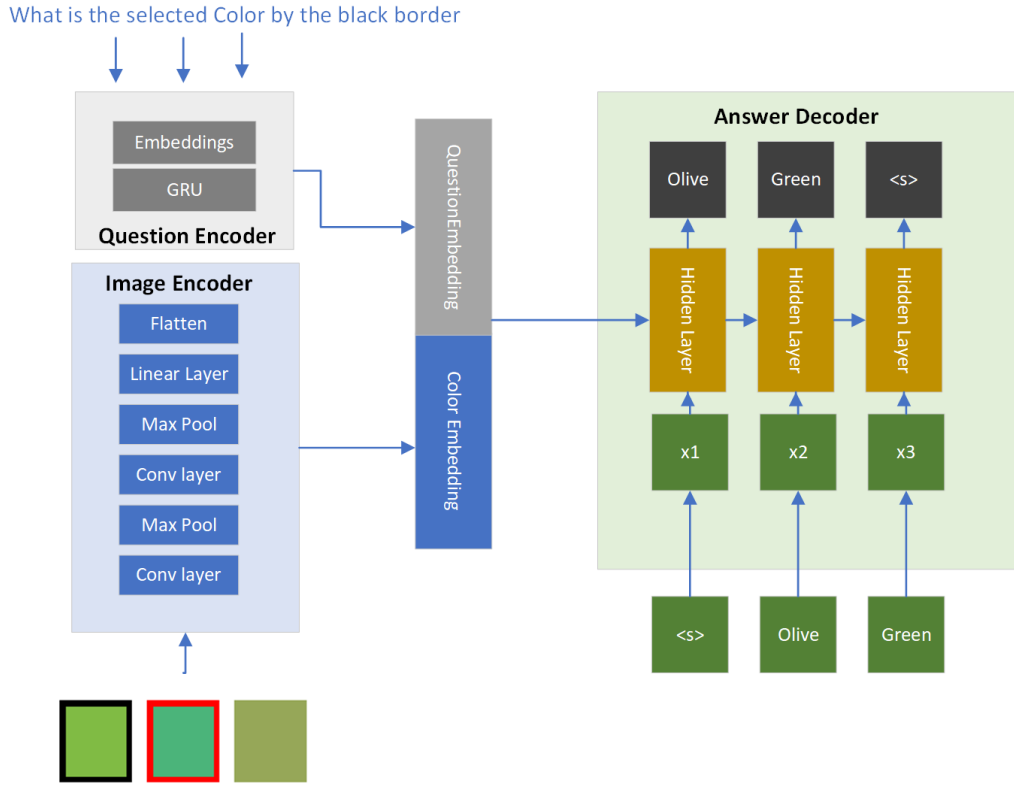


Figure 1: The Model uses a GRU based RNN network to encode the questions, this combined with image encoding, created by passing the image through a simple convnet network. The combined embedding is then passed into hidden layer of the Answer Decoder, where it decoded as part of another RNN based GRU network

5 Experiment

The model¹ is trained using the synthesized dataset described previously. Inputs to the model are the images and the question, with the output being the answer. Data was loaded in 64 sized batch chunks and trained over 10epochs while optimizing against an Entropy Loss cost function. The model makes use of teacher-forcing when training the AnswerDecoder against the expected target output.

After training the model was evaluated against a held of evaluation dataset and metric provided by the (Monroe et al., 2017) project. The evaluation metric called Listener Accuracy, measures how effectively the model communicates the color between a speaker and a listener agent.

This metric is based on computing probability of the utterance given a color set. All possible selections of colors are generated and evaluated and we pick the combination with the highest probability. The final score is computed by counting the number of computed entries that match the target selection.

¹Source: https://github.com/mathaix/colors_vqa

6 Model Analysis

Our Model was able to get a Listener Accuracy score 0.85 without any specific tuning. This compares favorably to benchmarks set by models based on (Monroe et al., 2017) architecture which achieved a Listener Accuracy Score of 0.91.

The results where also verified visually through Adhoc queries against the model with satisfactory results.

This is a bit counter-intuitive considering how simple the model is. The Convolutions neural network was very shallow and no specialized embeddings where used in the language model.

The good results could be rationalized as follows:

1. A reason for the good performance of the model is the data augmentation that happened during the synthesis of the VQA datasets. The model got to see multiple views of the same data and so tuned better.
2. Considering the simplicity of the underlying images, a deeper Convolutions Network is

quite unnecessary. A couple of different layers are sufficient to understand the features of this dataset.

7 Conclusion and Future Work

In this work, a new VQA dataset has been created and a baseline has been established against this.

A take away from this work is that there are two ways to train a model. One is to develop a custom model around the data by making use of specific feature engineering constructs and the other approach is to shape the data to fit a more generic model architecture. We have shown that this second approach works quite well. This seems intuitive when contrasting with how the human brain works and how the human brain is elastic enough to learn from a variety of different tasks.

This work can be taken further in a number of different ways.

1. Benchmark this data-set against established VQA models such as the (Tan and Bansal, 2019) model. .
2. Enhance the dataset to introduce more variety in the questions, since the current question set is pretty static.
3. More detailed analysis on how "condition" affects accuracy of the model.

8 Acknowledgements

This work was completed as participant in XCS224U: Natural Language Understanding professional education course through the Stanford Center for Professional Development.

Would like to thank Prof. Christopher Potts, on whose work this builds up on and for his guidance.

Would also like to thank the course coordinator Steve Haraguchi and course facilitator Madhulima Pandey for their timely help during the course of the program.

References

- Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. 2015. [Vqa: Visual question answering](#).
- Drew A. Hudson and Christopher D. Manning. 2019. [GQA: a new dataset for compositional question answering over real-world images](#). *CoRR*, abs/1902.09506.
- Lukasz Kaiser, Aidan N. Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. 2017. [One model to learn them all](#).
- Will Monroe, Robert X. D. Hawkins, Noah D. Goodman, and Christopher Potts. 2017. [Colors in context: A pragmatic neural model for grounded language understanding](#).
- Hao Tan and Mohit Bansal. 2019. [LXMERT: learning cross-modality encoder representations from transformers](#). *CoRR*, abs/1908.07490.