Abstract

Introduction

1 Dataset

There are two datasets that are directly related to understanding images with a help of dialog with an AI agent and are publicly available viz., *Guesswhat?!* And *VisDial*.

The first dataset *Guesswhat?!* was proposed in 2017 to train an AI system to hold conversation with a human. It was a two-agent cooperative game where the questioner had to guess the image which the oracle had selected in his mind by asking a series of questions. The dataset contains more than 150k games with over 800k question-answer pairs, but the dataset is restricted to answering only as *yes*/*no*/ *NA*.

The second dataset *Visdial* was also proposed in the same year and this is the main focus of the survey paper. Visdial evolved over multiple versions from v0.5 to v0.9 to v1.0.  All models discussed in this paper cater to either Visdial v0.9 or Visdial v1.0.

*Table 1 Visdial dataset splits [v1.0 above, v0.9 below]*

|  |  |
| --- | --- |
| **Training Dataset** | **Validation Dataset** |
| 1,23,387 images \* 10 rounds  (v0.9 train + v0.9 val) | 2064 images \* 10 rounds  Dense answer annotations |
| **Training Dataset** | **Validation Dataset** |
| 82,783  images \* 10 rounds | 40,504 images |

The dataset contains a dialog of 10 question-answer pairs (also called utterances) for each image. In v0.9, there are 1,232,870 images in total (taken from MS-COCO) [quote ref] dataset. In v1.0, the dataset evolved to include images from Flickr dataset to contain more than 130k images. The split of the datasets for train, test and val for the 2 datasets is mentioned in the table [1].

**Dataset analysis:**

MS-COCO dataset was used even for Visual Question Answering task but there are plenty of differences with VQA dataset. The dataset has more descriptive and longer answers than VQA dataset. Mean-length of the answers in VisDial is 2.9 words whereas it is only 1.1 words in VQA dataset. There are more unique answers in VisDial dataset. The top 1000 answers span only ~63% of the answers in VisDial, as against ~83% in VQA dataset. There are 10 question-answers per image in VisDial, as against only 3 question-answers per image. Definitely, the quality of the questions is much improved over the VQA dataset.

Another significant difference which VisDial has, unlike VQA and other related datasets like Visual7W, VisualGenome, FM-IQA, DAQUAR, COCO-QA dataset, is that the questioner was blind-folded. Intuition behind this type of data collection was to reduce the *visual priming bias* while asking questions. When the image is visible, often the questioner becomes biased to ask questions [11, 12] like “do you see <particular object> in the image” and thus overall the dataset becomes biased since a wild guess “yes” even without seeing the image would give higher accuracy. VisDial v0.9 is much better in this regard since it has roughly 47-53 distribution for yes-no answers, as against roughly 61-26 distribution for yes-no answers in VQA dataset.

VisDial v0.9 also performs better in terms of temporal continuity, which means that the entire dialog has fewer topics and every topic has more questions, mimicking the high continuity in human conversations. The two datasets were compared for 40 images and the scores [lower the better] turned out to be 2.14 ±0.05 topics for VisDial vs 2.53 ±0.09 topics for VQA for 3 successive questions.

2 Task & Evaluation Methods

2.1 Generative models

Given image I, dialog history {(Q1, A1), (Q2, A2), …., (Qt-1, At-1)}, follow-up question Qt, predict answer At.

2.2 Discriminative models

In addition to above given items, there is a pool of candidate answers. Goal of the discriminative model is to sort this pool of candidate answers, with topmost answer being the closest to the ground truth.

The pool of candidate answers consists of a ground-truth answer and 99 negative samples. This set of 99 answers consists of 50 plausible answers, 30 popular answers, 19 random answers. 1) Plausible answers are the answers to questions starting with similar tri-grams and bi-grams. To generate this set of answers, the GloVe embeddings of the first three words of all the questions and GloVe embeddings of all the remaining words in all the questions were concatenated. Further, the Euclidean distances indicated the similarity with the given question. 2) Popular answers are the most frequently output responses to the questions in the entire dataset. 3) Random answers justify their category name.

Idea of the inclusion of 99 negative samples is to fool the discriminative model to output the wrong answer. A robust discriminative model should be able to rank the ground truth higher in the returned sorting.

2.3 Evaluation Methods

Since the goal is to evaluate an open-ended free-form answer, metrics like BLEU, METEOR, ROGUE etc. are not suitable for the task [13]. Therefore, the metrics like Recall@k, Mean Reciprocal Rank and Mean Rank are used for the evaluation for Visdial v0.9 and Normalized Discounted Cumulative Gain (NDCG) was introduced for evaluation of Visdial v1.0, it was supposed to be better metric when the val and test datasets are human annotated. NDCG over top K-ranked (where K is the number of answers marked as correct by at least one of the four human annotators) “is invariant to the order of options with identical relevance and to the order of options outside of the top K.” [14]

3 Approaches

The general approach in most of the models is to use combination of encoders and decoders. Encoder combines the three modalities (image, question, history) into a common vector space and decoder is then used to either output the correct answer or return the sorted candidate answers, depending upon whether the model is generative or discriminative.

Most of the approaches for solving the Visual Dialog task are centered around proposing more efficient encoder modules, a bunch of them I will describe in the following section. Idea of the encoder module is to somehow combine the three modalities (image, question, dialog history) as efficiently as possible to form a feature vector. Most of the encoding strategies are based on incorporating the attention mechanisms. Attention mechanisms have recently gained a lot of focus in Visual Question Answering task [quote some papers] as well. Attention over image means you want to focus on particular regions of an image, which are most relevant to answering the question. Similarly, attention over history means you want to look up the most relevant questions that would make important contribution in answering the current question. For example, if there are multiple male persons in the image doing different activities and if the question asked is ‘what is he doing’, attention over dialog history will make sense to the question as to which person is being talked about in the previous questions. Similarly, attention over follow-up question can help to identify the objects and concepts the agent is asking about the image and accordingly you can identify the similar things from an image.

The first step of any encoder module is extraction of features from the image using a pre-trained CNN model. In the past, different architectures have been used like ResNet 101, VGG-19 etc. [quote if possible] for training the CNN models. Also, features from question and dialog history are extracted using LSTM-RNN model in most of the models. LSTM [quote] models are especially helpful in capturing the long-range dependencies and this can be particularly useful when we want to answer a question based on the dialog history (which can go up to 10 question-answer pairs). Different encoders have different strategies of feature extraction from question and dialog history. Some use separate LSTM models for every question in the dialog history and some pass the entire history through a single LSTM. The latter approach is more common and has given better results. [Give explanation if possible]

Apart from improving the encoder modules, some models have used Adversarial Learning based approach for improving the accuracy of the models, while some other approaches have been centered around Reinforcement learning and some are just the combination of the two.

Generative Adversarial Networks (GANs) have been extended to generate fake answers that are close to the ground truth, rather than generating fake images mimicking the real ones. GANs have two parts: Generator module (G) and Discriminator module (D). G tries to generate the answers as close to the ground truth as possible and D takes this answer as input and using some strategy or by computing a loss function, rewards G so that G can update its parameters and thus learn to generate more similar answers.

Reinforcement learning is also used in some of the models. Again, the setting remains similar to GAN-inspired models. Only difference is that in the former models, D is optimized for maximizing the reward so as to provide more informative feedback to G.

3.1. Visual Dialog

3.1.1 Encoders

Authors propose four types of encoders.

**Late Fusion (LF) Encoder:**In this encoder, the image is passed through a pre-trained CNN, dialog history is through an LSTM and the follow-up question through another LSTM to extract the features. The three feature vectors are then concatenated and further reduced to a 512-d vector space.

**Hierarchical Recurrent Encoder (HRE):** Two things are worth mentioning here. First, each question in the dialog history is passed through a separate LSTM unlike LF Encoder. Second, authors use LSTM models in two-level fashion. Lower-level (question-level) LSTM is used to capture the information of words of the follow-up question and the higher-level (dialog level) LSTM is used to capture the information of an utterance (pair of question-answer) in the entire dialog history.

Image features (I) are learnt by passing it through CNN (pre-trained on VGG-16).  I and Qt are then passed through an LSTM to give (I + Qt) vector. This output is then concatenated with the output from LSTMs (Ht-1) used for encoding history. This is repeated for every question in the dataset and finally, all such outputs (I + Qi where i =1, …, 10) are then passed through dialog-level LSTM to get the desired feature vector.

**HRE-with Attention (HREA):**This variant of HRE accounts for attending to a relevant utterance in the dialog history to answer a question. Before concatenating the output (I + Qt) with (Ht-1), (I + Qt) is passed through attention module to refer to a particular question(s) in the history.

**Memory Network (MN) Encoder:**[Read [9]] The basic intuition behind this is to again use attention mechanism to refer to a relevant question in the history. [Check out the difference with HRE-A except for hierarchy]. Similar to HRE, the utterances are passed through separate LSTMs and the outputs can be seen as *facts* in the memory bank. Question (Qt) is passed through another separate LSTM. Dot product is then computed between the question feature vector and each utterance in the history to compute attention-over-history probabilities. Finally, the resultant *convex* vector is passed through fc-layer of VGG-16 and then added to the (I + Qt) feature vector to produce the desired embedding.

3.1.2 Decoders

As discussed previously also, decoders can be generative or discriminative.

3.2 Best of Both Worlds: Transferring Knowledge from Discriminative Learning to a Generative Visual Dialog Model

3.2.1 Motivation

Generative neural network models trained with maximum likelihood estimation tend to produce very generic responses like I don’t know, I can’t tell. On the other hand, although discriminative models perform better than generative models (Das et al.), but they can’t be used for having real conversations with the chatbot. Discriminative models just return a sorting of the pool of the candidate answers in decreasing order of their relevance. So, the motivation was to combine the practical usefulness of the generative models and the good performance of the discriminative models.

3.2.2 History-Conditioned Image Attentive Encoder (HCIAE)

This is slightly improved over the encoders proposed by Das et al. Not just image and question are used to attend to the history, furthermore the attended history and question are used to attend to the image. This helps to focus on relevant regions in the image while answering the question.

3.2.3 Algorithm

The model proposed by the authors is much similar to Generative Adversarial Networks (GANs) [Goodfellow et al.] with only slight variations. Generative part (G) tries to generate an answer closest to ground truth whereas discriminative part (D) is used to assess the closeness of the answer and to give feedback to G.

Output from HCIAE encoder is fed to G to produce a distribution over candidate answers. An answer is then sampled from the distribution using Gumbel-Softmax [reference] sampler. This answer is then fed to D. Based on the input of 100 candidate answers including the ground truth answer, D pre-learns a function which tells the closeness of the generated answer to the ground-truth. The novelty of the paper also lies in using multi-class N-pair loss function [43] instead of multi-class logistic loss function. This loss function particularly prevents the correct but different answers from getting overly penalized. Finally, G updates its parameters based on the feedback from D.

3.2.4 Experiments & Discussion

3.3 Are you Talking to Me? Reasoned Visual Dialog Generation through Adversarial Learning

3.3.1 Motivation

This is another model inspired by GANs and applied to multimodal features. Additionally, reinforcement learning is used to update G. The difference from the previous model (section 2.1) is that D doesn’t take candidate answers as inputs and borrows the attention weights from G. These attention weights can be seen as a form of *reasoning* for the answer and thus, help in producing human-like responses (which are longer and descriptive).

3.3.2 Sequential Co-attention Encoder

We have three modalities – image, question and the dialog history. Authors use two of the modalities at a time to co-attend to the third modality. The difference with the other encoders is that CNN is pre-trained on Very Deep Convolutional Networks [6] and they complete a cycle of attention sequences (co-attention used three times sequentially instead of just one or two).

3.3.3 Algorithm

Just like a GAN, we have G and D. G takes in the *encoded* feature vector and passes it through LSTM (decoder) to generate an answer. D takes three inputs i.e. image feature vector, history feature vector and combined feature vector of follow up question and the generated answer. Output from D is the probability whether the dialog is human or not. This probability is used as reward to update G. D uses REINFORCE algorithm [3] to maximize the reward. Authors propose two more variants. First, they use Monte Carlo search to provide appropriate reward to the words of the generated answer, instead of just rewarding the entire sentence. Second, they use human feedback (teacher-forcing) [4][5] for updating G.

3.3.4. Experiments and Discussion

3.4 Stacked Co-Attention for VisDial1.0

3.4.1 Motivation

Although the prior works [Das et al., Lu et al., Wu et al.] have used attention models, but still they fail to answer questions related to fine-grained details in an image. Stacked attention models [7] work by first locating all the objects and concepts referred in the question and then pinpointing to the most relevant region in the image over multiple iterations. This type of multi-step reasoning helps resolve the problem of co-reference ambiguity. Also, authors use N-pair discriminative loss function which was shown to perform better than naïve logistic loss function [Best of Both Worlds].

3.4.2 Encoder

This encoder is an improved version of Co-Attention Encoder [Are you Talking to Me]. It is like a stack of three co-attention encoders.

3.4.3 Algorithm

Image features are extracted in bottom-up fashion using Faster RCNN (ResNet 101) pretrained on COCO-dataset, in a similar fashion with [8]. The rest of the architecture is pretty much similar except for the new encoder and discriminative loss function.

3.4.4 Experiments and Discussion

3.5 Image-Question-Answer Synergistic Network for Visual Dialog

3.5.1 Motivation

Other visual dialog models simply ignore the candidate pool of answers. Authors of this paper propose that even the candidate pool of answers can be used to attend to the image features and the dialog history.

3.5.2 Algorithm

Features for the image are extracted using Bottom-up and Top-down approach [Anderson et al.]. The model has two stages. 1) In the *primary* stage, feature vectors of image, dialog history and the question are passed through the Co-attention Encoder. The candidate pool of answers At is ranked using the N-pair loss function, similar to [Best of Both Worlds]. The spotlight of this stage is that the top N-ranked answers are selected and then appended with their corresponding question to form a new candidate set Bt. Idea is these N answers form *hard* samples, which means they are closest to the ground truth and other answers are called *easy* samples. According to the authors analysis, nearly 90% of the answers constitute easy samples and are actually not relevant for the question. 2) In the synergistic stage, feature vectors for Bt, dialog history and the image passed through the encoder and finally, the answers are re-ranked using a decoder.

Another interesting part of the paper was the fusion method they used for combining the different features. Borrowing the idea from [10], they used multi-modal bilinear pooling which fuses the feature vectors from different modalities more efficiently.

3.5.4 Experiments and Discussion

4 Results & Discussion

5 Future Work

References

[3] R. J. Williams. Simple statistical gradient-following algo- rithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992. 4, 5

[4]A. M. Lamb, A. G. A. P. GOYAL, Y. Zhang, S. Zhang, A. C. Courville, and Y. Bengio. Professor forcing: A new algo- rithm for training recurrent networks. In *Advances In Neural Information Processing Systems*, pages 4601–4609, 2016

[5]J. Li, W. Monroe, T. Shi, A. Ritter, and D. Jurafsky. Adver- sarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*, 2017.

[6]K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014

[7]Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked Attention Networks for Image Question Answering. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 21–29, 2016.

[8] Anderson et al.  2018 : P. Anderson , X. He, C. Buehler, D. Teney , M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image caption and vqa. In *CVPR 2018*.

[9] A. Bordes and J. Weston. Learning End-to-End Goal- Oriented Dialog. *arXiv preprint arXiv:1605.07683*, 2016.

[10] Z. Yu, J. Yu, J. Fan, and D. Tao. Multi-modal factorized bi- linear pooling with co-attention learning for visual question

answering.*IEEE International Conference on Computer Vision (ICCV)*, pages 1839–1848, 2017

[11]Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017.

[12] P. Zhang, Y. Goyal, D. Summers-Stay, D. Batra, and D. Parikh. Yin and Yang: Balancing and Answering Binary Visual Questions. In *CVPR*, 2016.

[13] C.-W.Liu,R.Lowe,I.V.Serban,M.Noseworthy,L.Charlin, and J. Pineau. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *EMNLP*, 2016.

[14] https://visualdialog.org/challenge/2019#evaluation