Visual Question Answering

Digbalay Bose [1] Nithin Rao Koluguri [1] Namrata Tammareddy [2] Aditya Mate [2]

[1]Department of Electrical and Computer Engineering,University of Southern California [2]Department of Computer Science, University of Southern California

Introduction

Visual Question Answering is one of the challenging AI tasks, which involves both reasoning about the image content and understanding the question to provide open ended natural language answer. Here in this work, we explore different deep neural network based models for generating natural language based answers. We aim to develop baseline models along with modifications of the current state of the art VQA setups and evaluate its performance on the standard VQA dataset.

Figure 1. Sample examples from the VQA dataset (The image, associated question and the answers are listed)

Dataset & PreProcessing

We consider the dataset given in `https://visualqa.org` for both binary question images and all class images. The details regarding number of images, questions and answers are listed as below:

Mode	Training	Validation	Testing
Images	82783	40504	81434
Questions	248349	121512	244302
Answers	2483490	1215120	-

Table 1. Table showing the number of training,validation and test images for real scenes

1. Fixed the number of question tokens being fed to LSTM/GRU to 14
2. Based on number of occurences of answers, created **two subsets** of the total dataset(2 classes i.e. yes/no and most frequent 1000 classes)
3. Pre-computed **resnet152** + **bottom-up R-CNN features** and stored them in hdf5 files for increased training speed
4. Pre-computed **BERT** [1] features for each question in train and validation set.

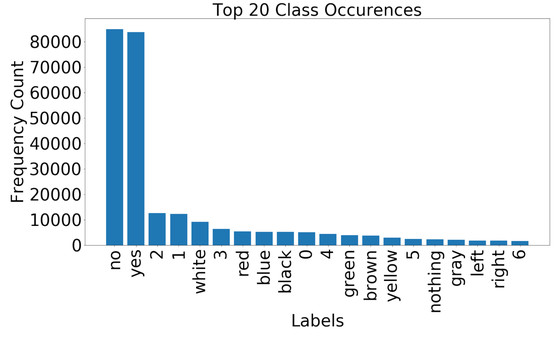

Figure 2. Histogram of Class Occurrences

Methods

1. VQA baseline architecture

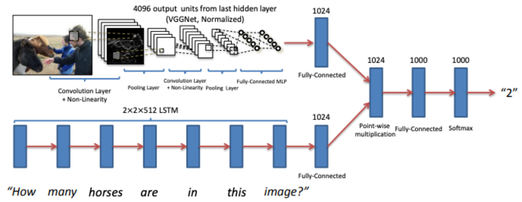

Figure 3. Baseline Architecture based on LSTM based question representation and its multiplication with image embedding

2. Question guided image attention and Multi-modal Factorized Bilinear Pooling

Question guided attention

$$\mathbf{a_i} = \mathbf{w_a f_a}([\mathbf{v_i}, \mathbf{q}])$$
$$\alpha = \mathbf{softmax}(\mathbf{a})$$
$$\hat{\mathbf{v}} = \sum_{\mathbf{i=1}}^{\mathbf{K}} \alpha_\mathbf{i} \mathbf{v_i}$$

α are the attention weights

Multi modal Factorized Bilinear Pooling

$$\mathbf{z_i} = \mathbf{x^T W_i^T y} = \mathbf{x^T U_i V_i^T y} = \sum_{\mathbf{d=1}}^{\mathbf{k}} \mathbf{x^T u_d v_d^T y}$$
$$= \mathbb{1}^T (U_i^T x \circ V_i^T y)$$

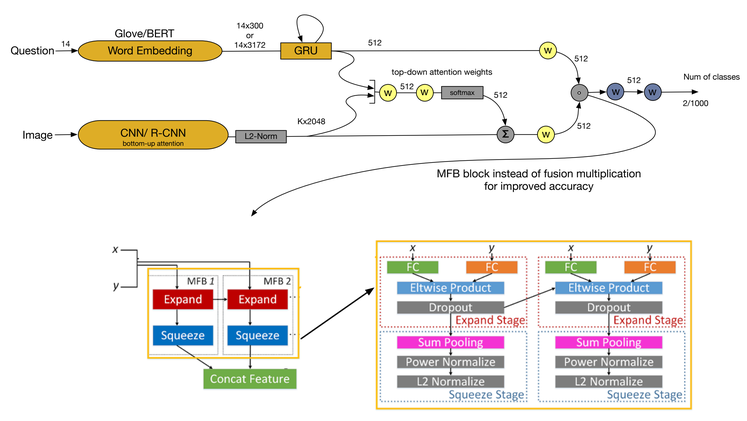

Figure 4. Modified attention architecture. Two main variants are considered: **Question guided attention + multi class classifier + multiplicative fusion**. **Question guided attention + multi class classifier + MFB fusion**

- VQA problem treated as a **multi-class classification problem(answers as labels)**.
- **Question guided attention** [2] for the image region features followed by multiplication based fusion of question/image embeddings and fully connected layers for multi class classification.
- **Question guided attention** [2] for the image region features followed by **Multi modal factorized bilinear pooling** [3] of question/image embeddings and fully connected layers for multi class classification.
- Our code repository is posted at `https://github.com/nithinraok/VisualQuestion_VQA`.

Experiments and Results

- *Configuration Details:* **Programming Framework:** PyTorch, **Batch size:** 512, **Optimizer:** Adam, **Learning rate:** 0.001, **GPU:** 2 NVIDIA GeForce GTX 1080 Ti, 1 K80, **Activation:** Leaky ReIU
- *Accuracy Metric:* Listed in `https://visualqa.org/evaluation.html`

$$accuracy = min(\frac{\#humans\ that\ provided\ that\ answer}{3}, 1) \quad (1)$$

Models	Yes/No	Number	Other
Att_GloVe_MFB_R-CNN_Uni_GRU_hid_1280_1000	80.11	45.23	60.95
Att_GloVe_mult_Resnet152_Uni_GRU_hid_1024_1000	76.29	39.55	54.11
Att_GloVe_mult_Resnet152_Uni_GRU_hid_512_2	72.25	-	-
Att_GloVe_mult_Resnet152_Bi_GRU_hid_512_2	73.05	-	-
Att_BERT_MFB_R-CNN_hid_1024_2	74.44	-	-
Adelaide-Teney_ACRV_MSR_3129 [2]	80.07	42.87	55.81
Adelaide-Teney_ACRV_MSR_2017_VQA_winner_3129 [2]	85.18	48.99	60.80
Co-Attention_ICCV2017_MFB_GloVe_VisGenome_3129 [3]	84.10	39.10	58.40

Table 2. **Att**: Attention baseline, **MFB**: Multi-Modal Factorized Bilinear Pooling, **Resnet152**: Image features ($2048 \times 7 \times 7$) **R-CNN**: Bottom up Image features (36×2048), **mult** : Multiplicative fusion of image and question embeddings , **2** :2 classes , **1000** : 1000 classes, **Uni**: Unidirectional , **Bi**: Bidirectional, **GloVe**: GloVe embeddings, **GRU**: Gated Recurrent Unit, **BERT**: Bidirectional Encoder Representations from Transformers, **hid**: hidden state size of GRU

GradCAM Visualizations- Decoding Failures and Successes

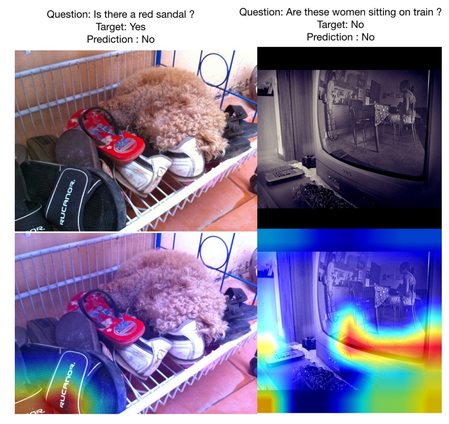

Figure 5. Examples from the VQA validation set showing correct and incorrect predictions with GradCAM. The model considered is the attention baseline with unidirectional GRU trained on resnet152 image features and GloVe Embeddings

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.
BERT: pre-training of deep bidirectional transformers for language understanding.
CoRR, abs/1810.04805, 2018.

[2] Damien Teney, Peter Anderson, Xiaodong He, and Anton van den Hengel.
Tips and tricks for visual question answering: Learnings from the 2017 challenge.
CoRR, abs/1708.02711, 2017.

[3] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao.
Multi-modal factorized bilinear pooling with co-attention learning for visual question answering.
In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.