Multi-Image Visual Question Answering

*Akhilesh Bhardwaj
*Harsh Raj
*Janhavi Dadhania
Supervisor : Prof. Prabuchandran K.J.
Indian Institute Of Technology Dharwad

Abstract

While a lot of work has been done on developing models to tackle the problem of Visual Question Answering, the ability of these models to relate the question to the image features still remain less explored.

We present an empirical study of different feature extraction methods with different loss functions. We propose New dataset for the task of Visual Question Answering with multiple image inputs having only one ground truth, and benchmark our results on them. Our final model utilising Resnet + RCNN image features and Bert embeddings, inspired from stacked attention network gives 39% word accuracy and 99% image accuracy on CLEVER+TinyImagenet dataset.

code: *https://github.com/harshraj22/vqa*

1. Introduction

Cross Modal task of visual question answering can be helpful in various areas for example visual aid to blind people, image captcha solver. Present state of the art methods solves the task in three parts, 1) image features extraction 2) Question feature extraction and 3) merging two modalities to finally predict the answer.

Good quality image and question features can improve the accuracy significantly. Few of the candidate models for image feature extraction are VGG [16], ResNet [8], RCNN [5]. LSTMs [9] are generally used to extract question features after encoding the question using word embeddings. Recently BERT [4] has proved to be an excellent text encoder which can further improve the performance. Merging of the two modalities is usually done by applying attention. Current state-of-the-art architectures are complex or tries to use different techniques like ensemble on top of simple structure for example oscar, pythia.

In this work we experiment with successful image feature extractors and question encoders which in our knowledge has not been tried yet. We also observe the significant improvement in the performance as we move from VGG to RCNNs and trainable embeddings to BERT.

We also propose modified VQA problem statement that first tries to find out the correct image that corresponds to the question and then answers the question.

Contributions.

- Empirical study of different feature extraction methods and their combination for feature representation.
- New balanced datasets.
- Experimental results with different training techniques.

2. Related works

Show, attend, ask and answer [13] model generalise all the previous works on VQA and claims that the simple architecture can give state-of-the-art performance. Stacked attention network [19] has the similar skeleton but it better utilize attention method. RstNet [21] proposes new method to ensure that spatial representation is not lost while flattening the image features. Oscar [15] uses new learning method using object tags detected in images to ease the model's learning. We use Stacked Attention Network model

---

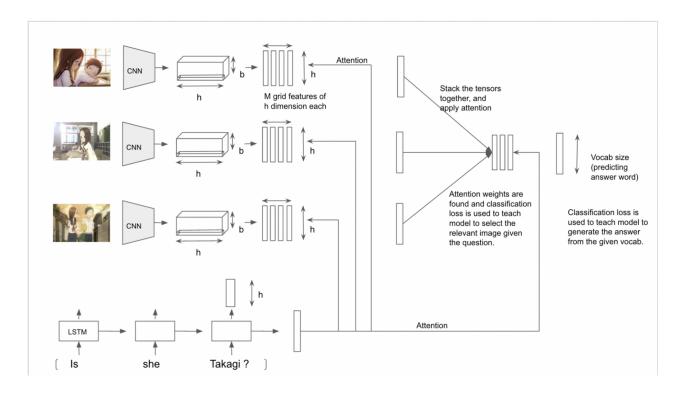*These authors contributed equally to this work

Figure 1. **Architecture –** A figure summarizing our base model

as our base model for below reasons,

1. Include general techniques used for VQA i.e CNN for image features and LSTM for question encoding and attention while merging.

2. Simple architecture gives satisfactory results comparable to complex ones.

There are many complex models (bottom up attention [1], ERNIE-VIL [20], Graph VQA [18], VL-BERT [17]) that show slight improvement in performance but also has significantly more number of parameters.

3. Method

Our proposed approach is derived from Stacked Attention Networks [19] which is for single image Visual Question Answering. We modified the model such a way that it can accommodate multiple images at a time and predict the answer along with the image from which it is most accurate to answer among the many input images. Model architecture is broadly divided into three parts.

- Image Encoder: We first extract features from images using pre-trained image classification models such as VGG, ResNet and object detection models like RCNN. For the image classification models, we consider the features just before the flattening step, while for the object detection models, we use the region proposals features. The extracted features are flattened and merged

to yield an output tensor of shape [-1,196,640].

- Question Encoder: We first tokenize the words in question to pass it through sequence model. We experiment with two different sequence models: LSTM and BERT. For LSTM, we use NLTK tokenizer while for BERT embedding we used BERT tokenizer because of their better compatibility with each other. After passing the questions through sequence model, we get question features as an output tensor of shape [-1, seq-len, 640] where is seq-len in max length of question that model can take.

- Attention: This part of the model works on efficiently combining the output of image and question encoder. We use multiple layers of Multi Head Attention [3] to merge the question features with the image features. The image features are used as queries and values, while the question features are used as keys.

Multiple input images are first passed through Image encoder while, questions are passed through Question Encoder. Lets say, for each data point in the dataset, a set of N images is input to Image encoder giving us output of [N, 196, 640] and meanwhile Question Encoder outputs [1, 30, 640] in this case seq-len is 30. Attention network takes input output of Image encoder and Question encoder and outputs weighted sum of the image features output

Method	CLEVER + Tiny Imagenet	VQA2.0
VGG	0.042	0.18
ResNet + Bert	0.32	-
Resnet + Bert + RCNN (Annealed)	0.39	-

Table 1. **Word Accuracy** of different models.

Method	CLEVER + Tiny Imagenet
VGG	0.24
ResNet + Bert	0.99
Resnet + Bert + RCNN (Annelead)	0.99

Table 2. **Image Accuracy** of different models.

Method	Without ImageLoss	With ImageLoss
VGG + trainable embeddings	0.038	0.042
ResNet + Bert + RCNN (Annealed)	0.32	0.39

Table 3. **Accuracies** with and without image loss.

of the Encoder. Then weighted sum image is used to answer the question in an old fashion single image visual Question Answering. Finally when Output in converted to answer vocab size using a Linear fully connected layer and converting output of Linear by passing it through softmax activation function and word with most softmax value is being output as an output of the whole model.

Loss Function: We treat Multi Image Visual Question Answering as a classification problem and use Cross Entropy loss for both Image as well as Word.

- Word loss + λ Image Loss: used a weighted average to two losses to train the network to predict the correct image to answer and then answer correctly after looking at the image.

- Word loss only: we may not always have liberty to have the correct image to train the model. Keeping this in mind we trained model with word loss only to try to train model to predict correct image using word loss only.

Annealed Training: In order to allow our model to learn the important correlations between the image features and the question features, we start the training with a high constant factor (λ) for image classification loss, and gradually decrease its value after each epoch. The model first learns to classify the correct image using the correct questions, and then to answer the question using the information from the image.

4. Experiments and Results

4.1. Datasets

We propose the following two datasets in order to benchmark our model's performance.

4.1.1 CLEVER + TinyImagenet

CLEVER [12] contains 100,000 synthetically generated images containing 3 to 10 objects with random shapes, color, size, materials etc and 999,968 questions. Tiny Imagenet [14] contains 100,000 images of 200 classes (500 for each class) downsized to 64×64 colored images.

We use CLEVER for the ground truth image, and questions, and randomly sample 3 images from Tiny Imagenet as non-ground truth image in our dataset.

4.1.2 VQA2.0 + Caltech256

VQA2.0 [6] is a balanced version of VQA [2] dataset containing 265,016 images and 5.4 questions on average per image. We use VQA2.0 for the questions, and detect all objects present in the image using YOLOv5 [11], and randomly sample images from Caltech256 dataset [7], which do not belong to the detected classes. This ensures that the 3 non-ground truth images are equally realistic and relevant to the ground truth image.

4.2. Results

We benchmark our model's performance on CLEVER [12] + TinyImagenet [14] dataset. Our base model with pre-trained VGG19 as backbone and trainable embeddings gives 4.2% word accuracy and 24% image classification accuracy on being trained with a weighted sum of Image Classification loss and word classification loss.

We then improve our model by adding ResNet as a better feature extractor, and Bert for pre-trained word embeddings, improving the word classification accuracy to 32% and image classification accuracy to 99%. Following Pythia 1.0 [10], we concatenate grid features extracted by ResNet and Region Features by RCNN as input to the model. This modified model is trained with annealed method giving 39% word accuracy and 99% image accuracy.

5. Conclusions

In this BTP we experimented with three different feature extractors i.e VGG, Resnet and RCNN and two different question encoder techniques i.e trainable embeddings and BERT, taking the stacked attention network as a base model. We experiment with different training methods and observe that,

1. Bert captures better features representations than training

our own embeddings from scratch

2. RCNN features provide improvements in image's feature representation than ResNet alone

3. Annealed Training helps in improving accuracy even further

6. Future Work

1. Benchmark all the above experiments on larger and comparatively more difficult VQA1.0 dataset that would help us understand advantages/disadvantages of multiVQA.

2. Improve the dataset further by using existing augmentation models like Neural Style Transfer.

3. Modify the loss function by using MSE loss for word embeddings instead of Cross-Entropy.

4. Further modify the task to predict multiple ground truth images.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering, 2018. 2

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015. 3

[3] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. Multi-head attention: Collaborate instead of concatenate. *CoRR*, abs/2006.16362, 2020. 2

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. 1

[5] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 1

[6] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[7] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset, 2007. 3

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[9] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 1

[10] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0.1: the winning entry to the vqa challenge 2018, 2018. 3

[11] Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, ChristopherSTAN, Liu Changyu, Laughing, tkianai, Adam Hogan, lorenzomammana, yxNONG, AlexWang1900, Laurentiu Diaconu, Marc, wanghaoyang0106, ml5ah, Doug, Francisco Ingham, Frederik, Guilhen, Hatovix, Jake Poznanski, Jiacong Fang, Lijun Yu, changyu98, Mingyu Wang, Naman Gupta, Osama Akhtar, PetrDvoracek, and Prashant Rai. ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements, Oct. 2020. 3

[12] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 3

[13] Vahid Kazemi and Ali Elqursh. Show, ask, attend, and answer: A strong baseline for visual question answering, 2017. 1

[14] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 3

[15] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks, 2020. 1

[16] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[17] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 2

[18] Damien Teney, Lingqiao Liu, and Anton van Den Hengel. Graph-structured representations for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2017. 2

[19] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering, 2016. 1, 2

[20] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934*, 1:12, 2020. 2

[21] Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, Yiyi Zhou, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Rstnet: Captioning with adaptive attention on visual and non-visual words. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15465–15474, 2021. 1