Visual Question Answering

Akanksha Shrimal
Indraprastha Institute of Information and Technology
MT20055
akanksha20055@iiitd.ac.in

Shivam Sharma
Indraprastha Institute of Information and Technology
MT20121
shivam20121@iiitd.ac.in

Vaibhav Goswami
Indraprastha Institute of Information and Technology
MT20018
vaibhav20018@iiitd.ac.in

1. PROBLEM STATEMENT

In recent times, there have been advancements in the field of Vision and Natural Language Processing.Visual Question Answering (VQA) is a challenging multi-modal learning task that helps us understand what our model truly sees and perceives from the image and how it responds to the question associated to the image. VQA requires an understanding of both visual and textual modalities simultaneously. Therefore, the approaches used to represent the images and questions in a fine-grained manner play key roles in the performance. The questions are text-based questions and are handled by Natural Language Processing while image features are extracted using sophisticated image processing techniques from vision domain.
We aim to build a deep learning system capable of answering open ended and yes no type questions on real world images. We used the VQA V2 dataset released by Virginia-Tech for our experiments that contains images from the MS-COCO dataset annotated with open ended questions and top answers.

2. LITERATURE SURVEY

VQA lies in the intersection of computer vision and natural language processing, which has attracted increasing interest from multiple research fields.Publically available datasets for VQA include DAQUAR, COCO-QA, VQA, Visual7W, and Visual Genome.
[1] used a simple but effective approach to solve VQA task. Text features extracted using a simple bag-of-words model and image features extracted using GoogleNet CNN model. Image and Text features were fused through point wise multiplication.

Currently, the mainstream VQA models are essentially based on attention mechanisms. Attention learns to attend to the most relevant regions of the input space and assigns different weights to different regions. [2] proposes a stacked attention networks (SANs) that allow multi-step reasoning for image.The SAN consists of three major components: (1) the image model, which uses a CNN to extract high level image representations (2) the question model, which uses a CNN or a LSTM to extract a semantic vector of the question and (3) the stacked attention model, which locates, via multi-step reasoning, the image regions that are relevant to the question for answer prediction. the SAN first uses the question vector to query the image vectors in the first visual attention layer, then combine the question vector and the retrieved image vectors to form a refined query vector to query the image vectors again in the second attention layer. The higher-level attention layer gives a sharper attention distribution focusing on the regions that are more relevant to the answer. Finally, image features from the highest attention layer are combined with the last query vector to predict the answer.
[3]also used stacked attention layers over images, along with pre-trained CNN model (ResNet) for image features and LSTM architecture for text-based features.
Most conventional visual attention mechanisms used in image captioning and VQA are of the top-down variety that give little consideration to how the image regions that are subject to attention are determined. [4] proposed a combination of bottom-up and top-down approach for Image captioning and VQA. The bottom-up attention puts forth the salient features and the top-down attention mechanism uses text-specific context to attend to the vital image regions and creates the model's feature weights. The VQA model is constructed using image features from Faster R-CNN which classifies localized objects and pretrained ResNet101 model and text features from GRU.
More recently, models based on co-attention have been widely used in VQA tasks, aiming to focus on salient regions of the image and critical words of the question.The co-attention mechanism can reduce unrelated information and obtain more meaningful features representations for image and question. [5] Proposed a Co-Attention Network With Question Type for VQA task. CNNs are pre-trained on ImageNet While Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) are used in extracting question feature. The image features and the question features are then fused via multi-modal pooling. They used Co-Attention module which includes self-attention based textual attention and question-guided visual attention to reduce noisy information.Question type module helped to reduce the search space by concatenating one-hot encoding of the question type directly to the multi-modal joint representation and thus producing accurate answers.
[6] took a new approach to handle the task as they created a dataset (Visual7W Dataset) which focused on 7W (what, where, when, who, why, how, which). The questions were categorized as *pointing* questions (what, where, when, who, why, how) and *telling* questions (*which*). They further used object-level groundings and then used attention on LSTM models to fetch better results. They used pre-trained CNN

model (VGG-16) and fed output of that as input to the attention based LSTM. The model achieved high accuracy along with detailed description of which kind of questions were better classified.

[7] boosted VQA by leveraging more powerful feature extraction techniques. They used Billinear Attention Networks (BAN) model, which uses Glove and GRU. They used Faster R-CNN based on ResNeXt for image feature detection and extracted features. For language, they replaced language model with BERT.

Present models focus on extracting high quality features from image using pre- trained models and text features using transformers.[8] used pre trained VisualBERT to extract image features. Using Transformes for text features citekhanMMFTBERTMultimodalFusion2020 outperformed many many state of the art models.

3. BASELINES

The dataset considered for evaluating the proposed VQA algorithms is given in https://visualqa.org/. The dataset contains open-ended questions about the images.
2 types of dataset are considered
a) **Yes/No Subset:** Yes/no subset consists of only questions with yes/no as answers.
b) **Top 1000 classes subset:** The top 1000 classes in terms of answer occurrences (labels) were considered while subsampling the entire dataset.

Dataset (train samples)		
Type	*Yes / No*	*1000 Classes*
questions	80789	187636
unique images	33793	39888
Dataset (val samples)		
Type	*Yes / No*	*1000 Classes*
questions	33289	76991
unique images	13843	16399
Dataset (test samples)		
Type	*Yes / No*	*1000 Classes*
questions	32329	74532
unique images	13501	15955

Table 1: no. of Samples (Train v/s Val v/s Test)

3.1 Baseline A

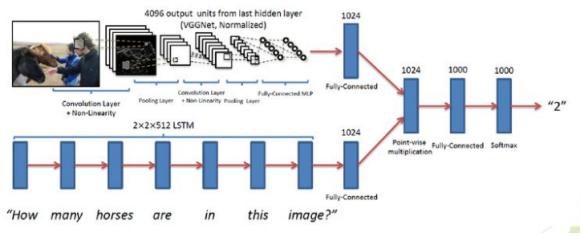

Figure 1: Baseline A Architecture - Point-Wise Multiplication

This model uses LSTM for textual questions and pre-trained CNN model (VGG-16) for image.
The questions are, first, integer encoded with Tokenizer and padded with zeros to have a uniform length of 24. The text is then encoded using 100-dimensional GloVe embedding. The encoded questions are fed to LSTM with 2 hidden layers to obtain 1024-dimension embedding for the question.
Image features are extracted using transfer learning from pre-trained VGG-16 CNN model, which is trained on a dataset of 14million images. The last layer (softmax) is removed and features from last hidden layer are used as image embedding of 4096 dimensions, which is connected to a fully connected layer with Leaky ReLU non linearity of 1024 dimensions so that to have question features and image features same shape.
The final combination of the textual features and image features is done by point-wise multiplication and then fed to a fully connected layer with Leaky ReLU non linearity and, finally, softmax layer to obtain probability of each of the 1000 answers.

3.2 Baseline B

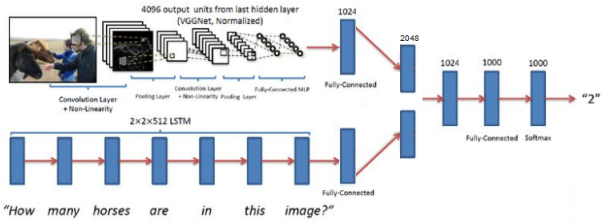

Figure 2: Baseline B Architecture - Concatenation

For this model, feature extractions techniques follow from Baseline A, i.e. LSTM with GloVe embedding for textual features and pre-trained VGG-16 for image features. Combination of the features obtained is done by concatenation resulting in a 2048 feature vector which is fed as input to a fully connected layer of 1024 dimensions. Finally, softmax layer is used to obtain probability of top 1000 answers.
If the classification type is Yes/No classification then an extra fully connected layer (1000,2) is used at the end to get the final probabilities of the two classes corresponding to the question.

Model Type:- 1000 Classes		
Baseline	*Train Accuracy%*	*Test Accuracy %*
A	62.0	33.0
B	27.0	26.0
Model Type:- Yes No		
Baseline	*Train Accuracy%*	*Test Accuracy %*
A	56.0	51.0
B	50.0	51.0

Table 2: Accuracy (Train v/s Test)

4. REFERENCES

[1] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus, "Simple Baseline for Visual Question Answering," *arXiv:1512.02167 [cs]*, Dec. 2015.

[2] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked Attention Networks for Image Question Answering," *arXiv:1511.02274 [cs]*, Jan. 2016.

[3] V. Kazemi and A. Elqursh, "Show, Ask, Attend, and Answer: A Strong Baseline For Visual Question Answering," *arXiv:1704.03162 [cs]*, Apr. 2017.

[4] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT: IEEE, Jun. 2018, pp. 6077–6086.

[5] C. Yang, M. Jiang, B. Jiang, W. Zhou, and K. Li, "Co-Attention Network With Question Type for Visual Question Answering," *IEEE Access*, vol. 7, pp. 40 771–40 781, 2019.

[6] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7W: Grounded Question Answering in Images," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 4995–5004.

[7] B. Liu, Z. Huang, Z. Zeng, Z. Chen, and J. Fu, "Learning Rich Image Region Representation for Visual Question Answering," *arXiv:1910.13077 [cs]*, Oct. 2019.

[8] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "VisualBERT: A Simple and Performant Baseline for Vision and Language," *arXiv:1908.03557 [cs]*, Aug. 2019.