Visual Question and Answering on Blur Images

Adarsh Pandey
PhD, IIITD
adarshp@iiitd.ac.in

Anwar Dilawar Shaikh
PhD, IIITD
anwars@iiitd.ac.in

Deepak Sharma
PhD, IIITD
deepaks@iiitd.ac.in

Sahil Kumar Singh
CSE 2020115, IIITD
sahil20115@iiitd.ac.in

K.Sibin
CSAM 2020307, IIITD
sibin20307@iiitd.ac.in

Aayush Singh
CSE 2020009, IIITD
aayush20009@iiitd.ac.in

March 13, 2023

1 Introduction

The problem addressed in this research paper is the need for Visual Question Answering (VQA) models capable of accurately processing and answering questions on blurry images. This issue is particularly concerning for visually impaired individuals who rely on VQA systems to provide information about their surroundings. The absence of VQA models that can handle blurry images can

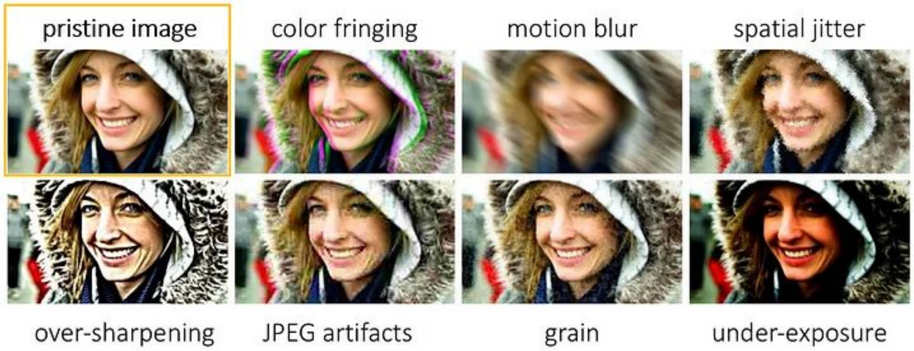


Figure 1: Caption

leave these individuals unable to get the assistance they need. Therefore, the research aims to address this gap by developing a better dataset and conducting a benchmark study to evaluate the performance of VQA models on blurry images.

Source: https://sh-tsang.medium.com/summary-m y-paper-reading-list-about-iqa-vqa-camera-tampering-blur-soiling-detection-327764516565

2 Motivation

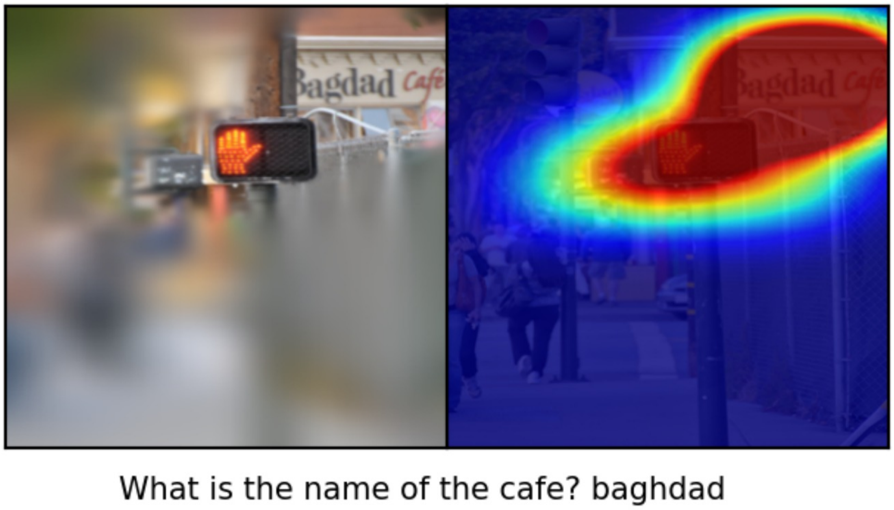
In many real-world scenarios, images may be degraded or blurry due to various factors such as motion blur, low lighting, or camera noise. By training and evaluating VQA models on blurred images, we can better simulate these real-world conditions and assess the model's performance in a more realistic setting.

3 Related work

There are multiple work on Blur detection/removal/classification. But not able to find any work on VQA on blur image. In a similar work(1), The researchers proposed a novel method to enhance person re-identification performance in real-world scenarios that are impacted by image degradation's. Their approach can handle real-world degradation's without needing large amounts of labeled data. They employed a degradation in-variance learning framework to extract robust identity representations for person re-identification. The method is self-supervised and disentangled, which enables it to capture and remove real-world degradation's without requiring additional labeled data. In another work(2), The authors introduce a new Image Quality Assessment (IQA) dataset based on a real-world use case. The dataset comprises 39,181 images captured by visually impaired individuals who used the VizWiz mobile phone application to learn about the images they took. The dataset provides a unique perspective on image quality issues as experienced by individuals with visual impairments (Chiu, Zhao, Gurari, 2019).

4 Proposed Work

To the best of our knowledge, no previous work directly focuses on visual question answering on blurred images. Therefore,our work is significant because it provides the benchmark for future studies. By creating a dataset and testing a VQA model on blurred images, we will establish a baseline for performance metrics and provided a reference point for future studies to build upon. In summary, this work fills a gap in the literature by addressing a previously unexplored area of visual question answering and providing a benchmark for evaluating the performance of VQA models on blurred images.



source : https://aclanthology.org/D16-1092.pdf

Figure 2: Caption

5 Evaluation

1) Accuracy: This metric measures the percentage of questions correctly answered by the VQA model.
2) Normalized Mutual Information (NMI): This metric evaluates the alignment between predicted and ground truth answers. It measures how much information is shared between the expected and ground truth answers.
3) F1-score: This metric measures the harmonic mean of precision and recall, calculated based on the predicted and ground truth answers.

6 Plan of Work

1) Assembling a varied and sizable dataset:

a)Identify and collect a large dataset of blurred images, illustrations, and inquiries that demand comprehension and thinking.

b)Ensure the dataset is diverse enough to cover various scenarios and situations.

2) Feature extraction using learned algorithms:

a)Select a pre-trained model such as VGG, ResNet, Detectron or Inception to extract features from the images.

b)Fine-tune the pre-trained model on the assembled dataset to improve its performance on the specific task.

3) Combining CNN and RNN for visual query assistance:

a)Use a CNN to classify the images and an RNN to analyze the query and generate the response.

b)Train the model end-to-end to optimize image classification and query answering performance.

4) Using attention mechanisms:

a)Implement attention mechanisms to enable the model to focus on specific parts of the image relevant to the query.

b)Use attention mechanisms to improve the model's accuracy and performance.

5) Data augmentation:

a) Apply various data augmentation techniques such as rotation, scaling, and flipping to increase the diversity of the dataset. b) Data augmentation can also help reduce over-fitting and improve the generalization capabilities of the model.

6) Regularization:

a) Apply regularization techniques such as dropout and weight decay to avoid over-fitting and improve the generalization capabilities of the model.

b) Regularization can help improve the model's accuracy and performance on new, unseen data.

7) Evaluate the performance of the model:

a) Test the trained model on a separate dataset to evaluate its performance.

b) If the performance is unsatisfactory, adjust the hyper parameters or retrain the model on a more extensive or diverse dataset.

7 Baseline:

Our problem statement was to find whether the image was recognized or not. For this, we have two Approaches. 1st was using ResNet-152 and 2nd using Detectron. We preferred to use Detectron over ResNet-152 as: -

1) **Model Architecture:** ResNet-152 uses specific deep learning model architecture for the image classification task, while Detectron uses various architectures for object detection and segmentation tasks.

2) **Object recognition vs Object Detection:** ResNet-152 is primarily used for image classification, where the task is to recognize the presence of objects in an image and assign them to a pre-defined set of categories. In contrast, Detectron is focused on object detection, which involves localizing and identifying objects in an image and segmenting them from the background.

3) **Training and interference:** ResNet-152 is typically trained on large scale image datasets, such as ImageNet, and then used for interference on new images. Detectron, on the other hand requires both training and interference,

as it needs to learn to detect and segment objects in new images-based examples provided during training.

Baseline Results:
Avg. precision - 80.0
Recall - 75.1
F1 score - 71.2

8 References:

1) Huang, Y., Zha, Z-J., Fu, X., Hong, R., Li, L. (2020). Real-world Person Re-Identification via Degradation Invariance Learning. IEEE Transactions on Circuits and Systems for Video Technology, 31(9), 3418-3428.

2)Chiu, T-Y., Zhao, Y., Gurari, D. (2019). Assessing Image Quality Issues for Real-World Problems. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 33-40)

3)Goyal, Y., Mohapatra, A., Parikh, D., Batra, D. (2016). Towards transparent ai systems: Interpreting visual question answering models. arXiv preprint arXiv:1608.08974.

4)Shih, K. J., Singh, S., Hoiem, D. (2016). Where to look: Focus regions for visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4613-4621).

5)Biten, A. F., Tito, R., Mafla, A., Gomez, L., Rusinol, M., Valveny, E., ... Karatzas, D. (2019). Scene text visual question answering. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 4291-4301).