A Visual Question Answering Model to Automate Nondestructive Evaluation Image Analysis

MEHRDAD S. DIZAJI and HODA AZARI

Abstract

This study introduces a Visual Question Answering (VQA) model designed specifically for nondestructive evaluation (NDE) applications. VQA models allow inspectors to interactively query NDE images—asking targeted questions like "Is there a crack?" or "Where is the defect located?"—and receive precise answers from the model. Leveraging deep learning and natural language processing, the developed system integrates image feature extraction (via a ResNet-50 model) and language generation capabilities (via GPT-2) to provide accurate, informative feedback. By enabling direct question-and-answer interactions, this VQA model significantly improves inspection efficiency, reduces potential errors, and enhances usability in practical field scenarios.

Keywords: Deep Learning, Visual Question Answering, nondestructive evaluation (NDE), GPT-2, language generation

1. Introduction

NDE techniques [1] plays a crucial role in assessing the structural condition of bridge elements. They can detect and locate hidden defects before they turn into serious problems [2]. detecting defects like cracks, voids, or other internal flaws early on can save money and, more importantly, save lives [3]. However, you need highly trained people to interpret the results correctly. That level of skill takes time to develop, and even then, human interpretation might be subjective. Mistakes can happen, and going through all the data manually can be slow and expensive [4]. That's where technology—and specifically artificial intelligence (AI) —starts to look really appealing.

Lately, there's been a lot of interest around using AI to assist with NDE. Tools like deep learning and natural language processing (NLP) [5] are changing the game by making it possible to automate some of the trickiest parts of the process. For example, deep learning models can be trained to recognize patterns and features in images that point to defects or abnormalities, even the subtle ones that might be easy to miss. On top of that, NLP can be used to automatically generate clear, human-readable explanations of what the AI is seeing, which is particularly useful for engineers and decision-makers [6]. This is precisely the objective of the NDE_Chat Platform. It's an AI-powered system to interpret NDE results more reliable without subjectivity. It uses state-of-the-art deep learning models—like convolutional neural networks (CNNs) to break down visual data and pull-out important features, and transformer-based language models to turn those insights into written descriptions. The end goal here is to reduce the workload on inspectors, cut down on human error, and give people more confidence in the results. By speeding up the analysis and making it more consistent, the platform can help bridge engineers catch issues earlier, understand material conditions better, and make smarter decisions when it comes to maintenance and repairs. It's all about

Mehrdad S. Dizaji, Ph.D., Post-Doctoral Research Fellow, Federal Highway Administration, Turner-Fairbank Highway Research Center, McLean, VA 22101, m.shafiei.dizaji.ctr@dot.gov
Hoda Azari, Ph.D., Nondestructive Evaluation Program and Laboratory Manager, Federal Highway Administration, Turner-Fairbank Highway Research Center, McLean, VA 22101, hoda.azari@dot.gov

combining the strengths of AI with the knowledge of experienced professionals to make data driven maintenance and preservation decisions.

2. Visual Question Answering (VQA) Model

Visual question answering is a task that requires answering questions about a scene based on an input image and natural language text. With the advent of deep learning, significant progress has been made in fundamental tasks such as image classification, motivating researchers to explore higher-level tasks that involve both vision and language, such as VQA. Early studies on VQA focused on extracting good features from the input image and fusing the two modalities to predict the correct answer. To extract features from the input question, existing natural language processing techniques have been adapted. The training process for the VQA model was conducted in two structured stages to enhance the model's multimodal reasoning capabilities. In the first stage, the model was trained on large-scale image–text pairs, which helped establish a foundational understanding of the relationships between visual content and corresponding textual descriptions. This phase allowed the model to learn generic visual-language representations such as spatial relationships, and contextual language associations. In the second stage, the model was fine-tuned on a domain-specific VQA dataset, carefully generated to include questions and answers explicitly grounded in visual content. The fine-tuning stage is critical for adapting the general-purpose vision-language model to the structured reasoning and answer-generation demands of the VQA task. VQA models are designed to interpret and answer questions about images by integrating techniques from computer vision and natural language processing. A typical VQA model architecture involves several key components:

1. **Image Feature Extraction**: Utilizes CNNs to process the input image and extract meaningful visual features.
2. **Question Processing**: Employs Transformer-based models to encode the input question into a feature representation.
3. **Feature Fusion**: Combines the visual and textual features through attention mechanisms, to focus on relevant parts of the image in the context of the question.
4. **Answer Prediction**: Generates a response based on the fused features, typically using a classifier to select the most appropriate answer.

2.1. Methodology

The goal is to teach the model to generate the correct answer yi to the question qi about the given NDE image Xi. Figure 1 shows the overall framework of our NDE VQA model, which is composed of an image encoder, a question encoder, and an answer decoder.

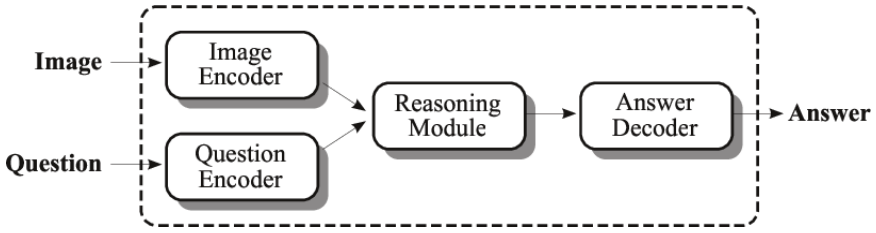

Figure 1. General architecture of the proposed medical VQA model.

The proposed model uses separate encoders for each input modality, followed by a shared decoder. A transformer-based visual encoder processes the input NDE image to extract visual features, while a

language encoder converts the input question into a corresponding language representation. Once both modalities are encoded, their feature vectors are combined to form a unified multimodal representation. This combined representation is then passed through multiple layers of the decoder to generate the final answer. The following section provides a detailed overview of the model architecture.

2.2. Evaluation Measures

BiLingual Evaluation Understudy [7] (BLEU) is automatic evaluation metric to measure the similarity of predicted answers and ground-truth by matching n-grams, as expressed below:

$$BLEU = BP \cdot exp(\sum_{n=1}^{N} w_n \log_e P_n),$$

where BP is the brevity penalty to penalize short answers, wn is the weight between 0 and 1 for logePn and $\sum Nn$=1wn=1, Pn is the geometric average of the modified n-gram precision, and N is the maximum length of n-grams. N-grams here are up to length 4.

2.3. Experiment Setup: Bridge Evaluation and Accelerated Structural Testing (BEAST)

BEAST is the full-scale bridge testing facility located at Rutgers University in New Jersey (Figure 2). The BEAST facility is the first facility nationwide capable of applying controlled and accelerated live load, environmental, and maintenance demands on full-scale bridge superstructures. The specimen is a multi-girder steel composite bridge (30 by 50 ft) with an 8-inch bare concrete deck and black rebar reinforcement. It will be subjected to rapid-cycling environmental changes and extreme traffic loading to speed up deterioration, as much as 30 times, in order to simulate 15-20 years of wear-and-tear in just a few months. The deck is supported by four I-beams as the main girders with one fixed and one open joint. At its initial design, the specimen is supposed to be exposed to over 8 million cycles of live loading (60 kips), 400 freeze-thaw and hot-dry cycles, as well as the application of deicing agents (6% brine solution) to simulate common winter maintenance practices. The BEAST database encompasses data collected on concrete bridge deck through five NDE technologies including Impact Echo (IE), Ground Penetrating Radar (GPR), Ultrasonic Surface Waves (USW), Electrical Resistivity (ER), and Half-Cell Potential (HCP). Data use in this study are the processed BEAST data that can be found on the https://infobridge.fhwa.dot.gov/ website. In this work only IE and GPR data is used to train the model. In the Table 1 number of samples to train, validation, and testing is shown.

Figure 2 .The overview of the BEAST facility.

2.4. Dataset Preparation

This research focuses on developing a VQA model, which enables users to interact with NDE data more dynamically by posing questions about specific aspects of the image and receiving intelligent, context-aware responses. The dataset preparation process for VQA builds upon the annotated image-caption pairs,

transforming them into structured question-answer formats. This involves curating a diverse set of domain-specific questions that address critical attributes of NDE images, including defect types, locations, dimensions, severity levels, and risk assessments. Expert-verified answers are then integrated into the dataset to ensure consistency and accuracy, creating a robust foundation for training an AI model capable of answering complex queries related to structural conditions. Table 1 shows the number of samples to train, validation, and testing.

Table 1. Number of samples to train, validation, and testing

		Number of samples		
Tasks	**Data**	**Train**	**Val**	**Test**
VQA	Image question answer tuples	550	35	10

2. Results & Discussion

The VQA model's performance was measured using the BLEU metric, and the results show that its answers closely match the correct, expert-provided ones. As seen in Table 2, the model scored 0.68 for BLEU-1, 0.65 for BLEU-2, 0.61 for BLEU-3, and 0.51 for BLEU-4—pretty solid numbers that show it's good at giving accurate and well-worded responses. For this test, the input was an Impact Echo image paired with a domain-specific question: "What type of defect is observed in the image?" The model correctly picked up on the low-intensity zones (the red-to-yellow areas with values between 1000 and 5000) near the surface, and it responded with an answer pointing to surface-level delamination (Figure 3). While its response ("PM") was a bit more concise than the expert-labeled answer, it still captured the key info needed to identify the defect. These BLEU scores—especially when compared to typical captioning tasks—suggest that fine-tuning the model on a specialized VQA dataset really helped sharpen its reasoning and response generation. Overall, the model shows it can understand detailed visual patterns and answer technical questions with accuracy, which is a big win for real-world use in non-destructive testing and structural evaluations.

Table 2. Evaluation of the trained model using BLEU Matrix

Matrix	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**
Model	0.68	0.65	0.61	0.51

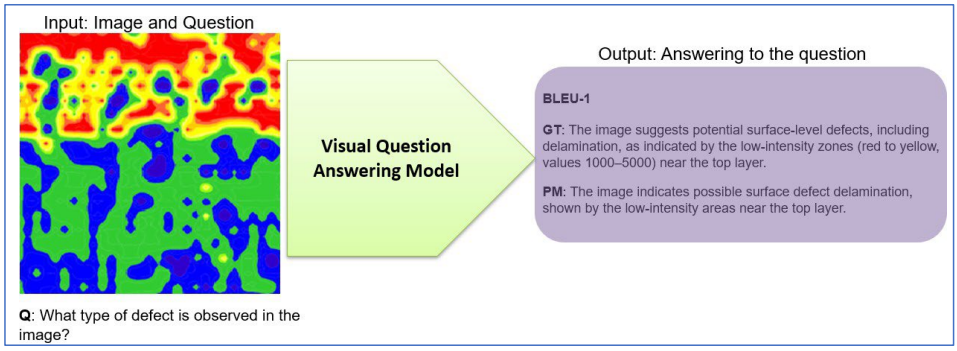

Figure 3. The input for VQA model is an Impact Echo image paired with a domain-specific question: "What type of defect is observed in the image?"

Continuing the VQA model evaluation using the BLEU metric, a second question was asked to see how well the model could pinpoint where a defect is located in the material just by looking at the image. The same Impact Echo image was used, but this time the question was: "Where is the defect located within the material or component?" (Figure 4). The model's answer was compared to expert-labeled ground truth, and again, the BLEU scores were strong—0.68 for BLEU-1, 0.65 for BLEU-2, 0.61 for BLEU-3, and 0.51 for BLEU-4. The expert answer pointed out that the defects were mainly near the top layer in low-intensity areas (values between 1000 and 5000), which suggests possible delamination or surface flaws. The model's response matched up well, noting that the defects were near the surface in those same low-intensity zones. Even though its answer was a bit shorter, it still captured the key technical points. This shows that the model is reliable across different types of VQA tasks, and that fine-tuning it on domain-specific questions really helped it learn how to reason about visual data. Along with the earlier example, this result proves that the VQA model can handle both identifying what kind of defect is present and figuring out where it is—making it a really useful tool for automated inspections in NDE applications.

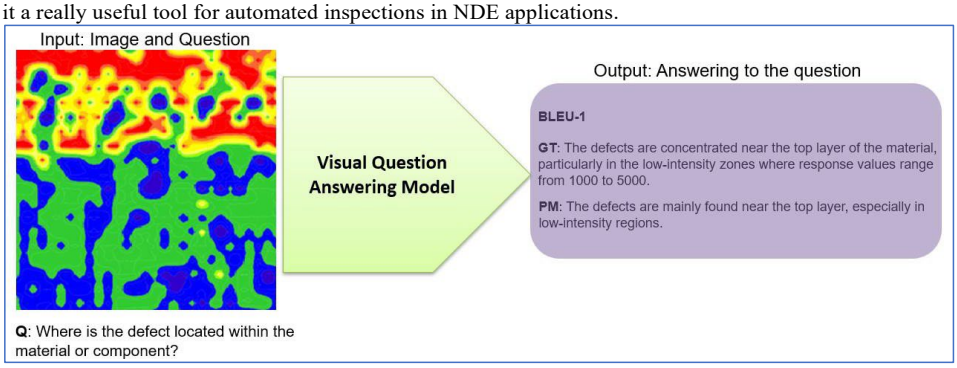

Figure 4. The input for VQA model is an Impact Echo image paired with a domain-specific question: "Where is the defect located within the material or component?"

The third VQA model evaluation looked at how well it could figure out the shape of a defect in an Impact Echo image. The question was: "What is the shape of the defect—linear, circular, irregular, or something else?" (Figure 5). This task went a step further than just spotting or identifying defects; it required the model to interpret the likely geometric form of what it was seeing. The BLEU scores stayed consistent with previous tests—0.68 for BLEU-1, 0.65 for BLEU-2, 0.61 for BLEU-3, and 0.51 for BLEU-4—showing that the model kept up its solid performance across different types of questions. The expert answer noted that, while the shape isn't clearly labeled in the image, defects like voids and delamination often show up in irregular or layered patterns. The model's answer was along the same lines, suggesting delamination usually appears in an irregular form—though it said it more briefly. This example points to a tricky part of VQA for NDE: questions that require inference or domain understanding (like guessing shape) can be abstract and not directly visible. Even so, the model gave a reasonable and accurate response that matched expert insights. Taken together with the previous results on defect type and location, the consistent BLEU scores show the model has solid potential for helping automate complex NDE image analysis.

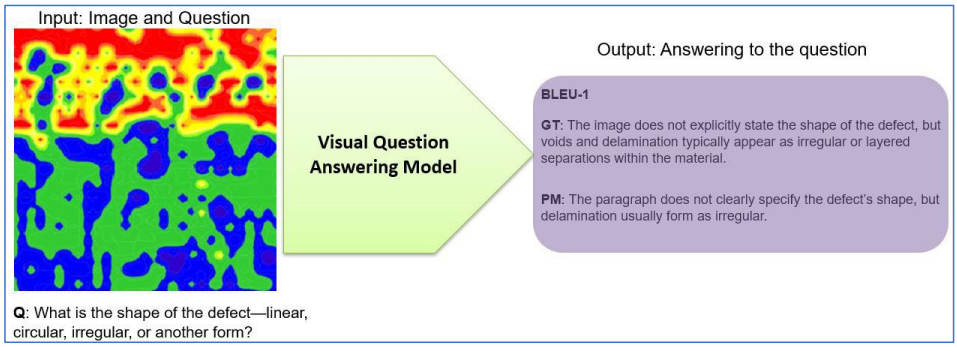

Figure 5. The input for VQA model is an Impact Echo image paired with a domain-specific question: "What is the shape of the defect—linear, circular, irregular, or something else?"

In the fourth VQA test, the model was asked to judge how **serious** the defect might be by answering: "How severe is the defect in terms of its potential impact on the material or component?" (Figure 6). This kind of question pushes the model beyond just describing or locating defects—it has to make a judgment call based on what the image shows and what that means for the structure's health. Once again, the model kept up its solid BLEU scores—0.68 for BLEU-1, 0.65 for BLEU-2, 0.61 for BLEU-3, and 0.51 for BLEU-4—showing its responses stayed consistent and well-formed. The expert answer explained that the cluster of low-intensity values near the surface could point to structural issues that might need closer inspection. The model picked up on the same idea, correctly flagging those surface-level, low-intensity zones as possible indicators of damage—though it explained things a bit more briefly. Even though the model's answer wasn't as detailed, it still nailed the key point about potential structural risk. Along with the earlier tasks (defect type, location, and shape), this example shows the model can handle tougher, more interpretive VQA questions. Its BLEU scores and alignment with expert insights suggest it's a strong tool for supporting automated structural evaluations in NDE work.

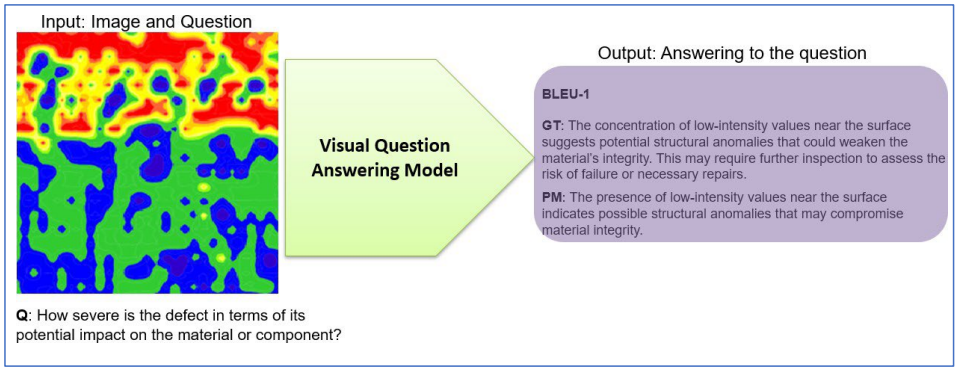

Figure 6. The input for VQA model is an Impact Echo image paired with a domain-specific question: "How severe is the defect in terms of its potential impact on the material or component?"

3. Conclusion

The ChatNDE-Figure-to-Caption framework's VQA model demonstrated strong results in interpreting NDE images. Trained and validated using the BEAST dataset, the VQA model effectively provided accurate and concise answers to targeted technical questions regarding defect types, locations, shapes, severity, and material conditions from Impact Echo data. Evaluated using the BLEU metric, the model consistently produced reliable responses, highlighting its effectiveness and practical utility in real-world inspection scenarios. Moving forward, enhancing the VQA model could involve expanding the dataset with diverse question types and scenarios, including multi-step reasoning and confidence-based answers, to further strengthen its applicability and performance in field conditions.

References

[1] L. Fülöp, M. Ferreira, A. Tuhti, and G. Rapaport, "Assessing the challenges of condition assessment of steel-concrete (SC) composite elements using NDE," *Case Stud. Constr. Mater.*, vol. 16, 2022, doi: 10.1016/j.cscm.2022.e00887.

[2] P. J. Shull, *Nondestructive evaluation: Theory, techniques, and applications*. 2016.

[3] J. Krautkra¨mer, H. Krautkra¨mer, and W. Sachse, "Ultrasonic Testing of Materials," *J. Appl. Mech.*, vol. 51, no. 1, 1984, doi: 10.1115/1.3167589.

[4] A. Osman, Y. Duan, and V. Kaftandjian, "Applied Artificial Intelligence in NDE," in *Handbook of Nondestructive Evaluation 4.0*, 2022. doi: 10.1007/978-3-030-73206-6_49.

[5] M. Treviso *et al.*, "Efficient Methods for Natural Language Processing: A Survey," *Trans. Assoc. Comput. Linguist.*, vol. 11, 2023, doi: 10.1162/tacl_a_00577.

[6] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.

[7] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2002.