Evaluation of a Visual Question Answering Architecture for Pedestrian Attribute Recognition*

Modesto Castrillón-Santana[1 [0000−0002−8673−2725]],
Elena Sánchez-Nielsen[2 [0000−0003−2114−4137]],
David Freire-Obregón[1 [0000−0003−2378−4277]],
Oliverio J. Santana[1 [0000−0001−7511−5783]],
Daniel Hernández-Sosa[1 [0000−0003−3022−7698]], and
Javier Lorenzo-Navarro[1 [0000−0002−2834−2067]]

[1] Universidad de Las Palmas de Gran Canaria, 35017
Las Palmas de Gran canaria, Spain
{modesto.castrillon,david.freire,oliverio.santana,daniel.hernandez,javier.lorenzo}@ulpgc.es
[2] Universidad de La Laguna, 38200
San Cristóbal de La Laguna, Spain
enielsen@ull.edu.es

Abstract. Pedestrian attribute recognition (PAR) ensures public safety and security. By automatically detecting attributes such as clothing color, accessories, and hairstyles, surveillance systems can provide valuable information for criminal investigations, aiding in identifying suspects based on their appearances. Additionally, in crowd management scenarios, PAR enables monitoring of specific groups, such as individuals wearing safety gear at construction sites or identifying potential threats in sensitive areas. Real-time attribute recognition enhances situational awareness and facilitates rapid response during emergencies, thereby contributing to public spaces' overall safety and security. This work proposes applying the BLIP-2 Visual Question Answering (VQA) framework to address the PAR problem. By employing Large Language Models (LLMs), we have achieved an accuracy rate of 92% in the private set. This combination of VQA and LLMs makes it possible to effectively analyze visual information and answer questions related to pedestrian attributes, improving the accuracy and performance of PAR systems.

Keywords: pedestrian attribute recognition · vision language models · Visual Question Answering.

* This work is partially funded by the the Spanish Ministry of Science and Innovation under project PID2021-122402OB-C22, TED2021-131019B-10, and by the ACIISI-Gobierno de Canarias and European FEDER funds under projects ProID2021010012, ULPGC Facilities Net, and Grant EIS 2021 04

1 Introduction

PAR is a field that encompasses interdisciplinary approaches to develop solutions for accurately identifying and understanding the attributes of pedestrians. This includes recognizing clothing color, accessories, and hairstyles to enhance situational awareness, manage crowds, and improve public safety and security. In PAR, specialized algorithms and models are traditionally customized to address the unique challenges associated with pedestrian attribute recognition. By leveraging advanced techniques from computer vision, pattern recognition, and machine learning, PAR aims to automate the analysis of pedestrian attributes, eliminating the need for manual intervention in tasks previously reliant on human intelligence.

In recent years, the rapid progress of Artificial Intelligence (AI) technologies, specifically deep learning applications, has led to significant advancements in PAR and garnered widespread recognition. These advancements have been facilitated by training deep neural networks on huge amounts of data and have revolutionized fields within the AI domain such as computer vision and natural language processing. Notably, the rise of LLMs has been exemplified by milestones like GPT-3. LLMs refer to AI systems pre-trained with vast amounts of textual data, in the order of hundreds of gigabytes or even terabytes of text data, showcasing unique language understanding, generation competence and the ability to perform multi-domain tasks without fine-tuning. Prominent LLMs, including GPT-3 [3], LaMDA [13], and LLaMA [15], have demonstrated remarkable capabilities in memorizing and utilizing extensive world knowledge. These LLMs exhibit emerging abilities such as in-context learning [3] and code generation [10]. Their capacity to harness and apply vast amounts of information represents significant advancements in the field. While LLMs have excelled in semantic tasks, their unimodal training strategy limits their extensive application with other data sources, such as sensors, cameras, and IoT devices. However, these data sources are crucial for comprehensive pedestrian attribute recognition, calling for innovative approaches to leverage the power of LLMs in PAR. Indeed, PAR is an important task in computer vision with numerous real-world applications. VQA, a prominent vision-language task, holds great potential in assisting various domains [1], including PAR. VQA allows pedestrian and traffic management centers to better understand their surroundings by providing answers to questions related to visual information. However, leveraging LLMs for VQA tasks can be challenging due to the inherent differences between visual and language inputs and the gap between language modeling and question answering. To overcome these challenges, a popular approach involves fine-tuning a vision encoder with a LLM [11]. This technique aligns the visual and linguistic representation spaces, enabling the model to accurately perform VQA tasks and establish the connection between visual and language information. By utilizing the pre-existing knowledge and generalization capabilities of the LLM, the model can answer questions about visual information without requiring specific training in the PAR domain.

This paper presents the iROC-ULPGC team's approach for the PAR Contest 2023 [7]. Our proposed pipeline leverages a pre-trained model without the need for additional training on the provided datasets. While using pre-trained vision language models for vision tasks is not a new concept, we can refer to the recent publication of the WISE Image Search Engine (WISE) [12]. This search tool utilizes a pre-trained vision language model called OpenCLIP, followed by a nearest neighbor search in the resulting high-dimensional feature space. The work by Sridhar et al. builds upon the achievements of Radford et al.[11], who demonstrated that deep models trained on large datasets containing millions of image-text pairs can effectively associate visual concepts with their textual descriptions. Hence, our original plan for the contest was to assess various vision language models. However, due to time constraints, we could only evaluate the performance of a fine-tuned BLIP-2 model for the VQA task. Despite this limitation, our contribution lies in adapting VQA techniques to the specific challenge of PAR. In various domains, including biometrics [5], zero-shot deep learning models provide the advantage of generalizability and adaptability to new tasks or domains without the need for explicit training data. In this regard, researchers have explored zero-shot VQA methods [8], which eliminate the need for ground-truth question-answer annotations. This approach enables the development of more generalizable VQA systems that can adapt to new questions and answer them accurately. Our findings confirm the impressive zero-shot image-to-text capabilities of the BLIP-2 model, yielding promising results with a mean accuracy surpassing 0.92 on the private set.

2 PAR Contest 2023

The PAR contest organizers provided the MIVIA PAR Dataset to participants [7]. This dataset comprises 105,244 images of cropped individuals, see Figure 2, separated into training (93,082) and validation (12,162) samples. Each sample is completely or partially annotated with numeric labels. The presence of a negative label for any sample refers to a non-annotated feature. The different features annotated are the following:

- Color of the upper and lower clothes. Two labels correspond to a single color associated with upper and lower-body clothes. Eleven possible colors are considered in the annotations: black (1), blue (2), brown (3), gray (4), green (5), orange (6), pink (7), purple (8), red (9), white (10), and yellow (11). The label in the brackets is associated with each color. Other colors are not considered in the dataset, neither are color combinations.
- Gender of the foreground person. The labels considered are male (0) and female (1).
- Bag presence. The labels considered are absence (0) and presence (1).
- Hat presence. The labels considered are absence (0) and presence (1).

More details about this dataset can be found in the description published by the contest organizers [7].

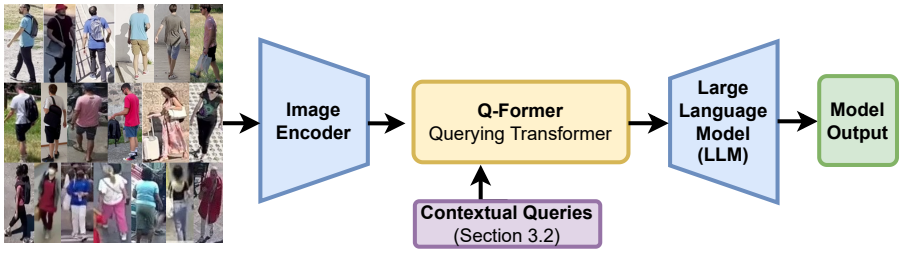

Fig. 1. The proposed pipeline for the PAR system. The devised process comprises three main modules: the image encoder, the querying transformer, and the large language model. In the first module, the image is encoded and passed to the second module, where related queries assist to extract the relevant features. The resulting tensor acts as an input to the LLM, completing the VQA process.

3 Proposal

3.1 Visual Question Answering

In recent years VQA has attracted the attention of the community, offering a meeting point for computer vision and natural language processing [2]. Unlike image captioning, where the image semantic information is extracted and expressed for humans, in VQA the information in the image is compared with a question or set of questions expressed in natural language. Among the set of applications identified by Barra et al. for VQA, surveillance and biometrics are valid real-world scenarios [14].

In our proposal, the adopted strategy uses a pre-trained BLIP-2 language model [9] trained on a large-scale corpus of text data and fine-tuned for VQA with the ViT base backbone [4], see Figure 1. The contribution of the BLIP-2 strategy is to leverage the training procedure. This is done in two bootstrapping stages: 1) the vision-language representation is learned from a frozen image encoder and 2) the vision-to-language generative model is learned from a frozen language model.

We have adopted a VQA approach because image captioning could not provide specific answers for the PAR Contest 2023 five subtasks. To illustrate this, the reader may launch the online demo[3] of the BLIP-2 image captioning model on the left sample depicted in Figure 2. We obtained the output '*A young boy is seen in this surveillance image*'. Below, we utilize a model trained with the VQA v2 dataset [6], which contains more than one million questions about COCO images.

The image captioning output may be helpful or enough for a general task but not for the particular subtasks requested in the PAR Contest 2023, where the proposals need to focus on the pedestrian's upper body and lower body colors, the gender, and the presence of bags and/or hats.

[3] https://huggingface.co/Salesforce/blip-image-captioning-base

Fig. 2. MIVIA validation set samples with: left) upper body annotation with a single color, with VQA reporting two colors, center) multiple individuals in the cropped area, and right) an individual with different jacket and shirt colors.

3.2 Contextual Queries

In PAR, incorporating contextual queries is essential when utilizing a VQA model. Contextual queries enable a deeper understanding of the visual scene and contribute to enhanced attribute recognition capabilities. By considering the surrounding environment, such as the presence of objects, landmarks, or social cues, the model gains access to additional contextual information that can provide valuable insights for attribute inference. Contextual queries allow the VQA model to go beyond analyzing individual pedestrian features and consider the broader context in which they appear. This holistic approach improves the model's ability to accurately identify and interpret various attributes related to pedestrians, facilitating more robust and comprehensive PAR results. By leveraging contextual queries, researchers can unlock the full potential of VQA models in addressing the challenges of pedestrian attribute recognition in complex real-world scenarios. After manually iterating with the validation set to increase the obtained accuracy, the final set of questions contained in the code provided to organizers is the following:

1. Is the person male or female?
2. What color is the person's shirt?
3. What color is the person's trousers?
4. Does the person wear a bag?
5. Does the person wear a hat?
6. Does the person wear a cap?

7. Does the person wear a jacket?
8. What color is the person's jacket?

The answers obtained from the model assign numerical labels to the evaluated image, explicitly targeting the resolution of five subtasks outlined in the contest: gender, upper color, lower color, bag, and hat. Certain answers directly correspond to specific labels. For instance, the response to question one provides the answer for the gender subtask. Similarly, a positive response to question four indicates the presence of a bag or similar item. In contrast, any positive response to questions five or six triggers a positive answer for the hat subtask.

Only the color subtasks required additional considerations within the scope of the study. The VQA model occasionally provides color responses that are not among the 11 colors used for annotation, or it may even provide combinations of colors. As an example, for the individual depicted in the left sample of Figure 2, the model's response was identified as *blue and white*. To address colors not originally included in the annotation, the validation set included alternative color options such as khaki, tan, plaid, and camouflage. For all such cases, a mapping to one of the 11 pre-defined colors was performed. In situations where a color not considered in the mapping appeared during the private evaluation, a random response was adopted. In cases where the model provided multiple color answers, the first color in the tuple appearing in the 11-color list was chosen as the mapping.

Considering these factors, the response to question three is mapped to the subtask of lower body color. However, for the upper body color, it was observed during evaluation on the validation set that the answer to question two alone was insufficient. This is because the VQA model may provide the color of the shirt, while the annotated color should refer to the jacket when one is being worn, as illustrated in the relevant sample depicted in Figure 2. To address this issue, a rule was devised by combining the responses to questions two, seven, and eight. This rule enables the determination of the appropriate color assignment for the individual based on the presence or absence of a jacket:

```
if person wears a jacket then
   color of upper body clothes = jacket color
else
   color of upper body clothes = shirt color
endif
```

4 Results

This section provides firstly a comprehensive summary of the results obtained from the validation set, which played a crucial role in determining the selection of questions for inclusion in the VQA procedure. Finally, the results provided by the organizers for the private set are also summarized.

Table 1. PAR 23 validation set results

Task	Acc.	Prec.	Rec.	F1
Upper color	0.805	0.805	0.805	0.801
Lower color	0.837	0.845	0.837	0.833
Gender	0.909	0.917	0.752	0.826
Bag	0.495	0.295	0.981	0.422
Hat	0.566	0.181	0.989	0.307

4.1 Validation set

The analysis of the validation set, comprising a total of 12,162 samples, has yielded encouraging results, as summarized in Table 4.1. We adopted *sklearn* to compute accuracy, precision, recall, and F1 score. For multiple classification problems, the weighted average is used, given the classes unbalance. The corresponding confusion matrices are shown in Figure 3.

Notably, these outcomes demonstrate promising performance across the 11 distinct categories encompassing the first two subtasks, specifically regarding color estimation. However, it is important to acknowledge that the human observer's perception of the jean's color in the images does not always align perfectly with the provided annotations, especially considering scenarios involving multiple individuals within the same image, as exemplified by the middle sample in Figure 3.

In terms of the binary subtasks, the obtained accuracy rates also display promising trends. However, it is crucial to highlight a couple of notable observations. First, the validation set exhibits an inherent imbalance between the number of males (8,674) and females (3,488), which necessitates careful consideration during analysis. Moreover, the accuracy rates for each class within this subtask exhibit noticeable variations, as evidenced in the corresponding confusion matrix in Figure 3.

Lastly, the evaluation of bag and hat presence in the validation set reveals a high recall rate, indicating successful identification of instances where these elements are present. However, the corresponding precision values do not reach equally high levels, implying a significant number of false positives. This observation, as evidenced by Table 4.1, suggests that the cropped area provided as input to the VQA model occasionally lacks sufficient contextual information to accurately determine the presence of these elements, particularly when they are positioned near the image boundaries. This limitation highlights the need for further investigations into methods that can better leverage contextual cues in such scenarios.

4.2 Private set

To ensure a rigorous evaluation process, the organizers of the PAR Contest 2023 have employed a mean accuracy metric that takes into account the performance across all five subtasks in a private set. For every subtask, given the number of

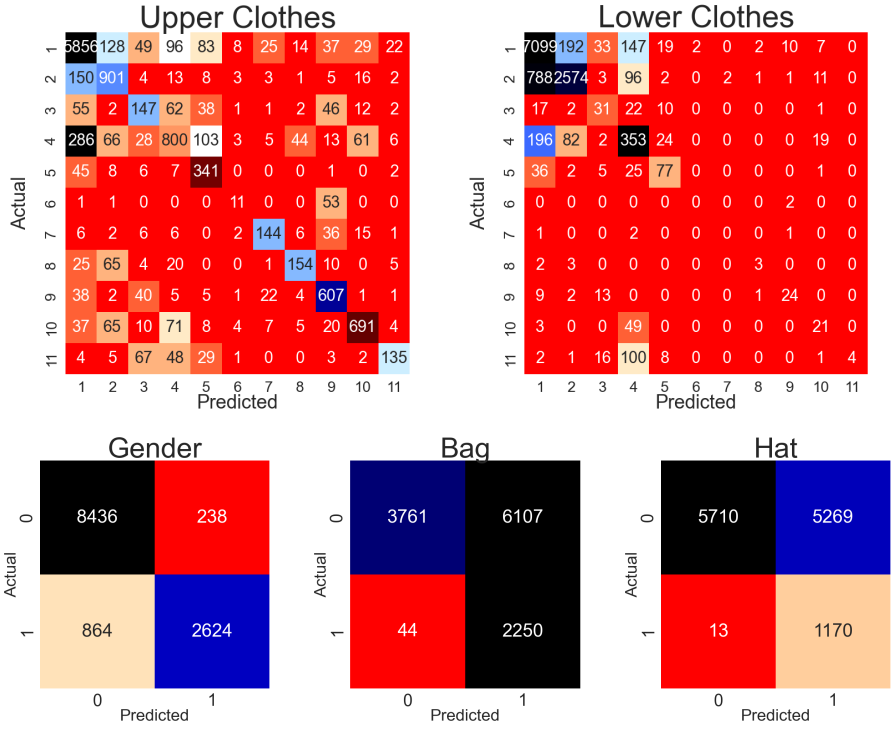

Fig. 3. MIVIA validation set confusion matrices.

samples K, the accuracy is computed by comparing the ground-truth labels of sample i-th, represented by g_i, with the corresponding predictions, denoted as p_i. This accuracy metric serves as a quantitative measure of the model's ability to accurately answer the questions.

$$A = \frac{\sum_{i=1}^{K}(p_i = g_i))}{K} \tag{1}$$

This subsection presents a summary of the results obtained from the private set, although specific details regarding the dimensions and distribution of classes across the various sets for the five subtasks are unavailable. Despite this limitation, the achieved results, as provided by the PAR Contest organizers and illustrated in Table 4.2, consistently demonstrate remarkably high rates of success for our proposed approach. It is crucial to emphasize that these outcomes were attained using a model that was not explicitly trained for the specific task at hand.

Table 2. PAR 23 private set accuracy results

Upper color	Lower color	Gender	Bag	Hat	Mean
0,9207	0,9081	0,9272	0,9215	0,9279	0,9211

5 Conclusions

Deep learning has brought computer vision forward by leaps and bounds. However, the optimization of huge networks with hundreds of layers results in complex mathematical models of millions of parameters whose inner workings cannot be easily understood by human beings. Without this understanding, it is very difficult for researchers to propose ways to improve their performance, hindering the progress in the field.

Multimodal networks, capable of achieving semantic understanding of images, represent a potential paradigm shift. Researchers no longer have to understand the inner workings of deep learning networks, instead they can concentrate on designing the image analysis strategy, planning the features to be examined and the decision making algorithm.

In this paper, we present a thorough evaluation of a VQA model based on the BLIP-2 architecture, specifically focusing on its performance within the context of the PAR Contest 2023. The obtained results, following a straightforward formulation of questions for the five subtasks under consideration, demonstrate highly promising outcomes. In the validation set, accuracies exceeding 95% were achieved in three tasks, while the color-related tasks exhibited an accuracy of 84%. Similarly, in the private set, all accuracies exceeded 90%.

These results reinforce the power of vision language models in addressing complex vision tasks and expand the realm of potential applications within the field, suggesting that we may be on the verge of very important changes in the way how computer vision problems will be tackled in the close future.

References

1. Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C.L., Parikh, D., Batra, D.: Vqa: Visual question answering. International Journal of Computer Vision **123**, 4–31 (2015)
2. Barra, S., Bisogni, C., De Marsico, M., Ricciardi, S.: Visual question answering: Which investigated applications? Pattern Recognition Letters **151**, 325–331 (2021). https://doi.org/https://doi.org/10.1016/j.patrec.2021.09.008, https://www.sciencedirect.com/science/article/pii/S0167865521003147
3. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T.J., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)

4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021), https://openreview.net/forum?id=YicbFdNTTy
5. Freire-Obregón, D., De Marsico, M., Barra, P., Lorenzo-Navarro, J., Castrillón-Santana, M.: Zero-shot ear cross-dataset transfer for person recognition on mobile devices. Pattern Recognition Letters **166**, 143–150 (2023)
6. Goyal, Y., Khot, T., Agrawal, A., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. Int. J. Comput. Vision **127**(4), 398–414 (apr 2019). https://doi.org/10.1007/s11263-018-1116-0, https://doi.org/10.1007/s11263-018-1116-0
7. Greco, A., Vento, B.: PAR Contest 2023: Pedestrian attributes recognition with multi-task learning. In: 20th International Conference on Computer Analysis of Images and Patterns: CAIP 2023. Springer (2023)
8. Kafle, K., Kanan, C.: An analysis of visual question answering algorithms. 2017 IEEE International Conference on Computer Vision (ICCV) pp. 1983–1991 (2017)
9. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models (01 2023). https://doi.org/10.48550/arXiv.2301.12597
10. Li, Y., Choi, D.H., Chung, J., Kushman, N., Schrittwieser, J., Leblond, R., Tom, Eccles, Keeling, J., Gimeno, F., Lago, A.D., Hubert, T., Choy, P., de, C., d'Autume, M., Babuschkin, I., Chen, X., Huang, P.S., Welbl, J., Gowal, S., Alexey, Cherepanov, Molloy, J., Mankowitz, D.J., Robson, E.S., Kohli, P., de, N., Freitas, Kavukcuoglu, K., Vinyals, O.: Competition-level code generation with alphacode. Science **378**, 1092 – 1097 (2022)
11. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021), http://proceedings.mlr.press/v139/radford21a.html
12. Sridhar, P., Lee, H., Dutta, A., Zisserman, A.: Wise image search engine (wise). In: Wiki Workshop (2023)
13. Thoppilan, R., Freitas, D.D., Hall, J., Shazeer, N.M., Kulshreshtha, A., Cheng, H.T., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H., Zheng, H., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., Chen, D., Xu, Y., Chen, Z., Roberts, A., Bosma, M., Zhou, Y., Chang, C.C., Krivokon, I.A., Rusch, W.J., Pickett, M., Meier-Hellstern, K.S., Morris, M.R., Doshi, T., Santos, R.D., Duke, T., Søraker, J.H., Zevenbergen, B., Prabhakaran, V., Díaz, M., Hutchinson, B., Olson, K., Molina, A., Hoffman-John, E., Lee, J., Aroyo, L., Rajakumar, R., Butryna, A., Lamm, M., Kuzmina, V.O., Fenton, J., Cohen, A., Bernstein, R., Kurzweil, R., Aguera-Arcas, B., Cui, C., Croak, M., hsin Chi, E.H., Le, Q.: Lamda: Language models for dialog applications. ArXiv **abs/2201.08239** (2022)
14. Toor, A.S., Wechsler, H., Nappi, M.: Biometric surveillance using visual question answering. Pattern Recognition Letters **126**, 111–118 (2019). https://doi.org/https://doi.org/10.1016/j.patrec.2018.02.013,

https://www.sciencedirect.com/science/article/pii/S0167865518300564, robustness, Security and Regulation Aspects in Current Biometric Systems
15. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models. ArXiv **abs/2302.13971** (2023)