POP-VQA – Privacy preserving, On-device, Personalized Visual Question Answering

Pragya Paramita Sahu * Abhishek Raut * Jagdish Singh Samant * Mahesh Gorijala

Vignesh Lakshminarayanan Pinaki Bhaskar
Samsung Research Institute
Bangalore, India

Abstract

The next generation of device smartness needs to go beyond being able to understand basic user commands. As our systems become more efficient, they need to be taught to understand user interactions and intents from all possible input modalities. This is where the recent advent of large scale multi-modal models can form the foundation for next-gen technologies. However, the true power of such interactive systems can only be realized with privacy conserving personalization. In this paper, we propose an on-device visual question answering system that generates personalized answers using on-device user knowledge graph. These systems have the potential to serve as a fundamental groundwork for the development of genuinely intelligent and tailored assistants, targeted specifically to the needs and preferences of each individual. We validate our model performance on both in-realm, public datasets and personal user data. Our results show consistent performance increase across both tasks, with an absolute improvement of ≈36% with KVQA data-set on 1-hop inferences and ≈6% improvement on user personal data. We also conduct and showcase user-study results to validate our hypothesis of the need and relevance of proposed system.

1. Introduction

In the last few weeks, large-scale multi-modal models like GPT-4 have taken over the world. With its seemingly, human-like understanding of images and its interaction with languages, these models seem ready to make us re-think how we do every day mundane tasks.

These powerful models, however, are battled by major challenges. The extremely large sizes (GPT-3.5 has approx. 175B parameters) along with high training requirements (350+ GPU days) remains a major bottleneck to make it device product compatible for widespread offline usage. The system ability to answer questions based on an input image frame (Visual Question Answering), is one such application that can not only empower users but also be an important, life-enhancing, service for any visually challenged user. From mobile-based accessibility, to IOT systems, such systems can allow an increased mode of intelligent interaction for users. Any such system, however, needs to be compatible to on-device implementations (low memory footprints) as well as have the power to generate personalized answers. Cloud based systems will invariably affect accessibility, especially across low income regions, and generalized systems are not of a lot of use in real-life daily scenarios. For example, in an IOT scenario, if the user wants to know who switched on the TV (while they are in a different room), they would ideally want the Smart Hub to respond with the name of the person rather than the generic answers of man or kid. Keeping this in mind, we propose an end-to-end, on-device system that generates personalized, user centric answers for queries on a visual frame using relevant meta-data information from on-device user knowledge graph(KG).

Figure 1. Example Scenario of systems powered by POP-VQA

We first build an on-device knowledge graph, centered

*These authors contributed equally to this work

around the user, using the information available from various sources like user profile, calendar, contacts and gallery meta-data. This knowledge graph stores information such as relationships of user, places visited, past and future events, occupation etc., in the form of triples. We enhanced our Vision Language model (VLM) based VQA system to generate personalized answers by empowering the model to choose and select relevant information from the user-centric knowledge graph, along with a deeper understanding of image and query alignment. This methodology allows our system to be trained on open-knowledge and to perform effective inference on user data. Proposed system significantly varies from existing SOTA techniques, in terms of task definition and training objective. Instead of relying on an external model to extract query-relevant entities from a KG, we train our model to choose the relevant information and generate answers. The complete system is then optimized to ensure efficient, on-device performance. We provide detailed results, on open as well as curated data-sets, to validate the efficiency and accuracy of the proposed system. We also provide extensive user survey results to corroborate the need of a personalized VQA system in user devices. To the best of our knowledge, this is the first work that creates an end-to-end application for personalized question answering based on an image input.

The rest of the paper is structured as follows: Section 2 talks about the related prior work. In Section 3 we look at the proposed approach, including details on knowledge graph creation process and proposed novel training methodology, followed by the description of all experiments in Section 4. Section 5 notes the environment and results of the conducted user study. We discuss possible impacts in Section 6 followed by conclusion and future work in Section 7.

2. Related Concepts

2.1. Vision Language Models

Vision language models (VLMs) leverage the synergy between visual and linguistic features to perform multi-modal tasks. While first proposed in the early 2010s, the recent advancements in data availability and computing powers, supported by powerful algorithms, has led to significant progress in the development of VLMs.

VLMs are trained on large parallel datasets of images and texts, which teaches it the relationships between these two modalities, via an aligned representation. This allows VLMs to support a multitude of generative and classifying tasks, such as image retrieval, image captioning, visual reasoning and visual question answering.

Most recent works [9, 10, 21, 28], use a cross-attention based transformer to achieve this. The transformer is given parallel image and text features and learns the alignment between each other. A tighter coupling in the aligned space is further realized by optimizing on various sub-tasks and corresponding loss functions. In general, the cross attention between language and vision features can be calculated as (language to vision and vision to language respectively):

$$h_i^k = CrossAtt_{L->R}(h_i^{k-1}, \{v_1^{k-1}, ..., v_m^{k-1}\}) \quad (1)$$

$$\hat{v}_i^k = CrossAtt_{R->L}(v_i^{k-1}, \{h_1^{k-1}, ..., h_m^{k-1}\}) \quad (2)$$

where for the layer k, the language features are $\hat{h}_i^k$ and vision features are $\hat{v}_i^k$. This cross attention is the "magic sauce" behind teaching models the feature correlation.

2.2. Visual Question Answering

Visual question answering (VQA) is a rapidly growing research area that enables machines to understand visual content and answer questions about it. With rapid growth in the fields of Computer Vision (CV), Natural Language Processing (NLP) and Knowledge Representation and Reasoning, VQA performance has also seen rapid growths. Earlier VQA models focused on combining the visual and text information through handcrafted features. With the arrival of Transformers [22] in 2017, along with significant growth in NLP techniques, VQA models saw marked improvements. Recent models consist of three main components: vision feature extractor, language feature extractor and cross modal attention layers. Image (vision features) and question (text features) are fed to cross modal attention layers to build an aligned understanding. LXMERT [21] gives fairly good accuracy than its predecessors using Faster RCNN for image features and BERT encodings for question features. Very recent models like COCA [27] and OFA [24] uses ViT [6, 22] features for image feature extraction.

A branch of VQA task has been growing recently known as **Knowledge aware VQA (K-VQA)**. These models work on methods of efficiently integrating open knowledge databases to effectively answer informative questions from images. Earlier works looked at methods of "question templating" to return external knowledge and form the answers. Recent works look at methods of injecting an external knowledge entity (returned from knowledge graph datasets) to allow the model to learn and answer [2, 3, 8, 19, 20]. The advent of Large Language Models (LLMs) haven't left this field untouched as well, with PICa [26] using GPT-3 to answer from descriptive captions for an image. All existing work, however, focuses on techniques of answering from open knowledge sources like Wikipedia entities. These works also limit themselves to only open knowledge domain, and neither provide results on general VQA performance nor do they extend their systems to personal knowledge integration.

2.3. Knowledge Graphs

Knowledge graphs are large-scale, multi-relational data structures built to effectively capture the relationships between multiple real world entities. Coined in 1972 by Edgar

W. Schenider [16] in the context of building modular instructional systems for courses, recent years have seen a huge impetus into building efficient large scale graphs that capture all common sense knowledge. This is achieved by extracting information from a variety of sources such as text corpora, databases, Wikipedia entries etc. Kertkeidkachorn et. Al [7] proposed a hybrid approach that combined traditional rule based methodologies with learned vector similarity for entity and relation extraction. More recently, Mondal et. Al [12] presented a completely deep learning based approach to construct graphs from unstructured data. Once created, SPARQL based techniques are used to query these RDF (Resource Description Framework) graphs and extract the relevant information.

A particular, and probably more relevant, extension of knowledge graphs is "Personal Knowledge Graphs (PKG)". Unlike traditional KGs, these are device and user dependent, and are built on top of all user related data that can be collated from various on-device sources (contacts, calendar, location etc.). This helps in building a structured graph resource about entities personally related to its user, their attributes and the relations between them. Most mobile solutions today build and use such personal KGs to help provide users with a personalized device experience.

For our work, we first construct such a PKG using information available on user's mobile device. This information is converted to "triples" [1, 14] and inserted in the knowledge graph. Further, some additional inferences are performed to deduce personal data such as, but not limited to, event-image associations, occasion, family and person identification etc.

We use the above PKG information, at a one-hop level, to power the model to generate relevant and personalized answers. For our specific purpose, accuracy can only be ascertained based on personal data, which by its definition is not publicly available. Hence, in order to compare our results with other SOTA Knowledge VQA based works, we include relevant common sense information from KVQA dataset [18] as meta-data during training and inference. This allows us to compare our performance on public datasets vis-à-vis other architectures (Section 4).

3. Proposed Approach

As described earlier, our aim is to build an end-to-end, on-device compatible privacy preserving model, that can be deployed on mobile and IOT devices to provide a seamless, interactive and personalized question answering experience for users.

Let us denote the VQA training data-set as $D = (I_i, Q_i, K_i, A_i)_{i=1}^{N}$, where I_i, Q_i, K_i and A_i denote the image, question, relevant knowledge and answer respectively of the i^{th} sample in a set of N samples. Meta-data knowledge can be detailed as $K_i = (E_k, R_k)_{k=1}^{M}$, where E_k represents the recognized entity and R_k denotes its relation to the primary user. Our training objective additionally targets to ground the answer A_i generation to the corresponding knowledge information K_i. Formally, the target is to maximize the conditional probability of the correct answer, given the inputs:

$$\max P(A_i \mid I_i, Q_i, K_i) \tag{3}$$

In the following sections, we describe the steps involved in building this system. Given the difference in task modalities between training and inference, we also provide detailed figures (Fig 2a and 2b, *next page*), describing the individual systems. The model is trained on a set of pre-processed, knowledge injected datasets (Section 3.3) that empowers the model to learn to generate image aligned answers for the user query and choose to "personalize" it based on injected external knowledge. At inference time, the model gets image information from user selection and external knowledge from the User KG. The model extends its learning to learn to pick and personalize the information to generate answers relevant to the user needs.

3.1. On-device, User-Centric Knowledge Graph

To merge together the world of personalization with visual question answering,enhance our answers with personal, user information. This graph is built using data from the User Profile, Calendar, Contact, Gallery meta-data (including location data) available on mobile devices. Further details on data sources and usage is described in Appendix B. KG construction primarily focuses on triples related to the following categories:

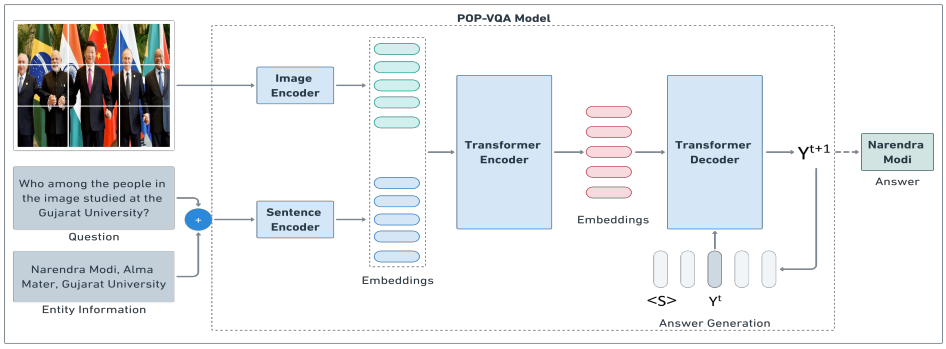
- **Relationships**: family, friend, colleague, boss, pet etc.
- **Events**: annual events like birthday, anniversary etc., life events like graduation, marriage etc. and daily events like exercise, meeting etc.
- **Locations**: places associated with user such as home, work, school etc.

In cases where the desired information is not available in the input data sources, we infer additional triples using image processing techniques and pre-defined inference rules. A few examples of such inferences are:

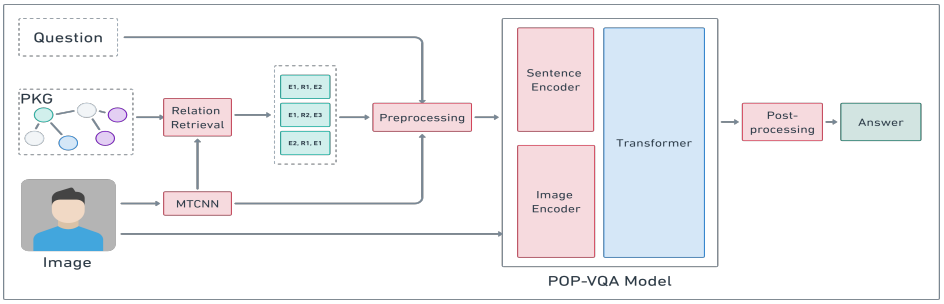
- Person and Person Name Inference
- Relationship with user
- Occasion of image
- Hero of event (e.g. for the event "daughter's birthday", hero is daughter)
- Association b/w events and images

These triples are stored on an on-device embedded graph database that is built on top of RDF4J [1] and made available to applications via a SPARQL query interface. For every chosen image, the external module returns a sub-graph that includes all relevant entities and its relation to the user.

[1] https://www.rdf4j.org



(a) Model Training with Open Knowledge entities



(b) On Device Model Inference with User Knowledge Entities

Figure 2. Diagrammatic representation of task definition and process during training and inference

3.2. Data-set Description and Curation

Most VQA models developed in recent years have been trained and evaluated on the popular VQA2.0 data-set [4]. This data-set provides a diverse collection of image, question and answer triplets. However, while the VQA2.0 data-set has been instrumental in training and bench-marking VQA models on good general understanding, it lacks any data for knowledge-based VQA (K-VQA). For this task, various independent data-sets have been introduced that incorporate the requirement of external knowledge to answer user queries. KVQA [18], OK-VQA [11], A-OKVQA [17] are a few examples of such openly available data-sets.

Our target objective however, doesn't directly align to either of the tasks. Our objective is more user-centered, to build an application that users can use in their real day-to-day lives. This model should, therefore, not only have high accuracies on general VQA but also be able to integrate user personal information and provide enhanced, personalized answer when needed. With this in mind, we curate our training and testing data-sets by a combination of VQA2.0 and KVQA data-sets (60 : 40 ratio of KVQA samples to VQA2.0 samples), to allow the model to learn a general alignment (from VQA2.0) as well as knowledge inference capabilities (from KVQA). Our experiments in Section 4.1 also validate the need for mixed data usage. While KVQA only contains open knowledge, our training methodology ensures that the model is efficiently able to mold itself to integrate any knowledge graph (user-centric for our application) at inference.

- ***VQA2.0 Data-set***: The VQA2.0 data comprises of 658k question and 121k images. The data-set covers a wide range of topics, image types and question types. The significance of the VQA2.0 data-set in our work stems from our objective of developing a VQA system that possesses a comprehensive understanding of the scene depicted in the image and can generate answer accordingly
- ***KVQA Data-set***: The KVQA data-set comprises of a collection of 184k question-answer pairs, centered around 18k distinct individuals with 24k images. The questions within the data-set requires sophisticated reasoning abilities of multiple entities and relations to answer. The information about the 18k entities can be readily found in the Wikidata. The KVQA paper proposes models that have been trained and evaluated using both closed and open world experiments.

Date of birth	Place of birth	Date of death	Place of death
Occupation	Gender	Religion	Linguistic Ability
Spouse	Alma Mater	Career Details	Current Organization

Table 1. List of chosen closed-world relations from KVQA data-set

As described earlier, we want to target only closed world experiments to closely align our model to the task of personal answer generation - grounded to the user on-device knowledge graph. Hence, for our experiments, we chose to use the 12 closed world entity relations (Table 1), to closely align the experiment with the context of our on-device VQA system We particularly make the choice of not including any multi-hop categories as the data-set required multiple levels of open knowledge facts (eg, "Who is the founder of the party that the person in the right belongs to?"). Within the scope of on-device knowledge, such information would never be available at the user level. However, we include multi-entity and multi-relation categories as they are more relevant to our target task.

3.3. Pre-processing

We face a unique situation with respect to modality differences during training and inference. During on-device inference, the user knowledge graph API returns all the related meta-data (at a 1-hop level) with respect to the chosen image. Such an interaction, while crucial at a system level where multiple applications access the same knowledge graph, makes the training process more challenging for us. Instead of having an external system that "selects" the correct answer from the complete graph [13,29], our system needs this "personal" knowledge integrated with each sample meta-data. Our model, thus, needs to be trained to not only understand the question and its relation with the image, but also needs to be taught how the meta-data provided relates to the selected image. For example, for a query on the lines of "Who is standing on the left of Mark", the system not only needs to be taught what "standing on left" represents, but also needs to learn the alignment in the location co-ordinate (Bbox) of Mark and Marie. Our pre-processer is thus, designed to merge the user query with the returned meta-data information into a combined space with a [SEP] token. To further refine performance over generative accuracies and keep token length in check, we implement the following enhancements:

- ***Spatial Information***: In keeping with the pre-training objective of OFA [24], we add the entity bounding box (BBox) information to the sample meta-data. During training and inference, these entity BBoxes are extracted using MTCNN [5]. We integrate this along with the entity name to teach the model spatial alignment to person name.

- ***Personal Entity Names*** : One issue we noticed during the generative process was the system's inability to correctly spell out more complex names (Jacobo Árbenz → Jacob Arbenz). To handle these errors and reduce the token length, we replace all names with placeholder information and re-map it to the required answer in post-processing.

- ***Entity Personal information***: We scan the KVQA data-set and choose the subset of examples most relevant to us (Table 1). For these samples, all the required information (birth details, occupation etc.) is tokenized and added to the meta data. This allows for task commonality between training and inference.

- ***Open Knowledge***: Certain questions in the training data, even at 1-hop level, needed external knowledge (such as the dates of World War II). While such samples are irrelevant for our inference, we wanted to include it to allow for comparisons with SOTA models. Such information is integrated with the question itself (open knowledge not needed for on-device inference as it is out of scope for our solution).

The VQA2.0 data-set doesn't include any open knowledge information. However, to ensure fair data representation, placeholder information is added in the meta-data for such training samples.

3.4. Model Description and Training Methodology

We build our POP-VQA model on top of a pre-trained VLM model. We choose OFA [24] as the foundation of our experimentation due to:

- OFA-base model reports a size of 180M parameters, making it more suitable for on-device optimizations than the recent extremely large scale models that go into billions of parameters.

- OFA builds task-agnostic capabilities by mapping various alignment tasks such as question-answer problems, making its understanding stronger.

- OFA's pre-training data-structure aligns with our meta-data injection methodology, making the model easier to adapt to external knowledge

OFA uses a transformer based architecture to build the aligned space on top of RESNET101 (for image feature extraction). They also use the techniques of sparse coding to reduce the sequence length of image representation. This model is pre-trained on 5 tasks (visual grounding, grounded caption generation, image text matching, image captioning and visual question answering) on ≈50M images. The model is optimized with cross entropy loss and uses beam search for effective generation. This is pre-trained for 300K steps with Adam Optimizer.

We use this pre-trained model to fine-tune for our task of personal knowledge based VQA. Training data consists of a combination of VQA2.0 and KVQA datasets, as described in Section 3.2. No personal data is used for the training procedure, as an important metric for our system

was its ability to effectively generalize to user personal data. We inject external knowledge (relevant to 1-hop from user) while training using the methodology described in Section 3.3. All images are resized to $480X480$. This model is then fine-tuned for $30K$ steps with a learning rate of $5e-5$ and label smoothing of 0.1. Exponential moving average with a decay of 0.999 is used to further refine and generalize the model. Such a training technique allows for a two-step alignment process. The text encoder self-attention allows the model to learn the relation between the query and the relevant meta-data information, while the multi-modal causal self-attention learns the relation from image to query and meta-data . This allows the system to generalize effectively, even when the task is modified from a generic knowledge based VQA to a more personalized, user specific VQA. The system is able to retain its visual reasoning knowledge, allowing for high performance on general visual queries as well.

3.5. On-Device Deployment

The model is implemented and trained using Pytorch library [15]. We follow the pruning method as proposed in [25] followed by int8 quantization to reduce the model size and make it suitable on on-device deployments. We are able to achieve a compression ratio of 80% by following this approach. The optimized model is then converted to ONNX format [2] for efficient inference on mobile devices.

4. Experiments & Results

We aim to evaluate our model on two crucial foundations - (i) Real-time, user personal data performance and (ii) Performance on open-datasets. This two pronged approach helps us validate that not only does our model learn the taught objectives, but also easily generalizes to out-of-scope, open situations with person-centric knowledge. Additionally, as we target a solution that can be integrated in mobile and IOT systems, performance validation on real user data becomes mandatory. We thus split our testing into two phases:

- ***General Testing***: We test our model on two benchmark data-sets - KVQA and VQA2.0. As described earlier, these data-sets are chosen for its closest relevance to our task. Our model is evaluated on the curated test-sets of both these samples (43K image-question pairs combined from KVQA test data and VQA2.0 validation data).

- ***Personal Testing***: With the aim of getting real-time validation, we choose a group of 100 independent users (age and gender demographics equally spread). For the purposes of evaluation, we collect their personal images (total of 5k questions on 1.5k images) with relevant user KG information. The system is also integrated on their personal devices, to allow further testing and relevant reporting of user experience. All participants are clearly explained the target of our model, and encouraged to ask questions that they would in their daily lives. More details are described in supplementary materials.

[2]https://github.com/microsoft/onnxruntime

Question Type	MemNet	UNITER	POP-VQA(K)	POP-VQA(K+V)
1-hop	61.0	65.7	**89.8**	83.7
Boolean	75.1	94.6	95.7	**97.8**
Comparison	50.5	90.4	89.6	**94.1**
Counting	49.5	**79.4**	73.2	75.0
Intersection	72.5	**79.4**	72.3	69.5
Multi-Entity	43.5	77.1	**94.9**	90.0
Multi-Relation	45.2	75.2	**93.27**	92.7
Spatial	48.1	21.2	**83.89**	68.6
Subtraction	**40.5**	34.4	37.0	26.7
Overall	–	–	**85.8**	83.5

Table 2. Comparison of model accuracy on KVQA test data-set based on question type. POP-VQA(K) refers to model trained only on knowledge data. POP-VQA(K+V) refers to model trained on both KVQA and VQA2.0 data-set. Overall accuracies not mentioned for SOTA models due to task mismatch.

Category	OFA-base	POP-VQA	Category	OFA-base	POP-VQA
Activity	65.99%	61.04%	Complex Features	71.38%	62.16%
Boolean	89.73%	75.78%	Complex Inference	69.24%	62.74%
Color	84.19%	72.54%	Object identification	85.62%	74.89%
Comparison	84.24%	73.68%	Person identification	37.30%	86%
Counting	69.10%	77.24%	Spatial understanding	69.52%	93.8%
Location identification	42.78%	97.01%	Miscellaneous	66.79%	73.49%
Overall	**75.77%**	**71.31%**			

Table 3. Comparison of POP-VQA (K+V) model on VQA2.0 curated data-set with OFA-base performance. Categories have been manually generated from the VQA2.0 mentioned question types

4.1. Model Performance : Knowledge VQA

Our first and most important performance metric remains the model performance on knowledge based VQA samples. As described in Section 3.2, we made the deliberate decision to focus on only 1-hop data. This choice was driven by the fact that, in our scenario, only the entity depicted in the image and its direct relationships with other entities within the image hold significance, and best aligns with the requirements of our personalized VQA system. We first fine-tune our model on this subset of 134K image-question-answer triplet and evaluate the performance on a similar subset of KVQA test data. We utilized the works of Garcia et.al in [2] and Vickers et.al in [23] as our baselines for performance evaluation. We choose these works primarily because of their closest similarity to our target applications. Other works are less relevant as they focus on analysis of external knowledge graph to answer questions instead of

training the model to build this understanding.

In Table 2, we note the performance of our model, with comparisons across the various chosen subsets. We saw a remarkable 36.7% improvement in accuracy for 1-hop data when compared to the [23]. We attribute this significant enhancement to the improved training of self-attention layers. Spatial reasoning also saw a huge accuracy jump of 74%. This progress can be credited to the incorporation of grounded captioning during the pre-training phase of the OFA model. An interesting observation is seen on subtraction question types (eg. calculating the age gap between two individuals). As can be noted, there is no significant improvement compared to earlier models. On analyzing the predictions we found out our model tends to predict a number that was relatively close (± 1) to the correct answer but not same. We aim to delve deeper into this issue and identify potential solutions, ultimately enhancing its accuracy and robustness.

We also evaluate our POP-VQA(K) model on the VQA 2.0 val data, with the assumption that pre-training on VQA, would allow more decent performance in general, non-knowledge based scenarios. However, our findings indicate that the results obtained were rather poor. These outcomes suggest that although the model demonstrated proficiency in KVQA tasks, its performance on the more extensive and diverse VQA2.0 data-set was not as successful - proof that the latent space had become aligned to a specific tasks. This highlights the need for further fine-tuning the model in order to enhance the generalized answering capabilities.

4.2. Model Performance : Integrated VQA

As described above, to enhance the system performance of generalized answering, we conduct fine-tuning and experimentation with a mixture of data-sets. We combine the data-sets described in Section 3.2, and empirically determine that a $60:40$ ratio of knowledge VQA to generalized VQA data samples ($134K$ and $90K$ samples respectively) provides the best performance. This aligns with our intuition as well, where while learning to understand an image to answer queries is difficult, the difference in modality for personalized VQA (as compared to pre-training) needs significant training samples. OFA pre-trains with the question answering task, making it easier for the model to retain that ability with fewer samples during fine-tuning. This data was then pre-processed (Section 3.3) and used to fine-tune the model. All hyper-parameters remain same as earlier. We note the detailed results and analysis in Table 2 [POP-VQA(K+V)] and Table 3. In Table 3, the category-wise performance is provided after collating the 63 mentioned question type in the data-set [4].

A quick comparison of results in Table 2 and Table 3 show that not only is our system now able to generate personalized answers, but it still retains the image understanding capabilities. Especially in questions that cover person and location identification, and counting, we note an increase in performance on both knowledge mixed data-set (Table 2) and VQA2.0 data-set (Table 3). This is a direct reflection of the commonality of tasks, and shows a strong alignment between the image representation to the meta-data information.

We also note slight drop in general VQA performance (Table 3) as compared to OFA-base model. Especially in cases of complex understanding and comparisons, POP-VQA under-performs with respect to OFA-base. This however, seems a direct casualty of the reduced VQA2.0 training samples to maintain the required sample ratios. However, as we discuss next, this degradation is not noted in real-time performance. This is also a reflection of the kind of questions that are asked in daily scenarios, which are more aligned to identification of the person, location and activities and their interaction with each other. When tasked with more generalized questions, aligned to daily life situations mobile, IOT and smart robot applications, POP-VQA outperforms a generic VQA model in all aspects.

In further work, we will include other open-knowledge datasets, accordingly giving us the power to increase the VQA2.0 training samples. This, we believe, would lead to more performance gains, especially in the cases where there is significant scope of improvement.

4.3. Model Robustness

Strong performance on open-datasets only prove model validity within the same data domain. While necessary, it is by no means sufficient to prove the robustness and real world applicability of a solution. We, thus, test our model on user's personal data post solution integration with their devices. We collect 1.5K images with 5K questions and its related meta-data. Model performance on this data-set is noted in Table 4. As we cannot compare performance with other SOTA models on this data-set (due to mismatch in task definitions), we instead compare our performance with a general, non-personalizing, model (OFA-base). We compare the core KPIs of accuracy and latency to get a measure of system performance. Accuracy@N is measured as the percentage of times the required answer came in the top-N options. Latency is the end-to-end system processing time on Samsung S22 device. Integration of personalization does increase system latency, due to increased data processing at inference time. However, this marginal increase ($\approx 90ms$) leads to significant accuracy improvements ($\approx 7\%$) as well as increased answer relevance. Especially in cases like person/gender identification, counting and location/event understanding, this additional meta-data grounds the output to a relevant domain and provides validation for generated answers - thus reducing erroneous cases. Further qualitative examples to validate increase in answer relevance has been

Model	Accuracy			Latency	Model Size
	Top-1	Top-3	Top-5		
Generic VQA	77.71%	89.79%	92.23%	682ms	220MB
POP-VQA	82.91%	93.57%	93.89%	770ms	

Table 4. Model Performance on personal data testing and on-device KPIs

provided in supplementary materials.

4.4. Model Efficiency

Another major metric for any on-device solution to be commercially viable is in terms of its memory-space and latency. As described in Section 1, on-device deployment was crucial for wide-spread and inclusive application. Along with low footprints, real-time answering is imperative for a seamless and interactive experience. We thus, evaluate our model on these parameters and note the performance in Table 4. Note that all latency numbers are provided as an average of inference runs on user devices.

5. User Study

In an effort to understand the true impact of the proposed solution, we conducted a user study with test participants. We integrate our system with user's personal devices and allow participants to ask the system any question from their personal gallery photos. The results of this study are noted in Table 5. As we can see, users not only found this system more accurate and relevant than a generic VQA system, but also believe that such solutions are the need-of-hour to make smart systems truly intuitive and interactive. Participants also noted and liked that with an on-device solution, their data is truly private but they can still interact with the systems at a personal level. This study further validates the need of personalization with generated answers.

Model	System Accuracy	Answer Relevance	Need for Solution	Ease of Usage
Generic VQA	8.1	2.6	N/A	N/A
POP-VQA	8.8	9.3	9.5	9.1

Table 5. User Study: Average scores (/10) from 100 participants

6. Societal Impact - The Good & the Bad

Our work is motivated with solving some of the various challenges that the recent, large-scale models create. As our model name suggests - we build our model to have 3 major features, namely (i) privacy conserving, (ii) low model sizes and (iii) personalized inferencing. The targeted aim of on-device compatibility manages to "kill 2 birds with 1 stone". On-device inferencing not only preserves user data privacy (as no information is moved out of device), but also provides users with limited internet access (especially in low resource locations), an equivalent experience. This allows for a more universal usage. Optimized model sizes also directly translates to a faster inference time, which coupled with personalized inferences creates a seamless, human-like interaction experience for users. This is especially useful for visually challenged users, empowering them to independently navigate and interact with the world. It goes a long way in promoting inclusion and autonomy, enhancing their overall quality of life.

However, no story is complete without some limitations. To meet our objective of an on-device, personalized solution, we had to trade-off with model capabilities. Keeping a product perspective, we limit our model's knowledge to a 1-hop, personal level. This curtails the scope of queries the system can answer. Our solution, is also limited by the accuracy and speed of knowledge graph creation. Using a pre-trained model on open-datasets also increases the chances of bias and skewed representation in the model. As we conceptualize this model for real-life implementation, having systems in place to keep the bias (gender/race/sexuality) in check, becomes more crucial.

7. Conclusion

In this work, we propose a personalized, on-device, VLM to answer any user query for an input image. While we show our experiments on top of a pre-trained OFA model, we believe our training methodology can be integrated with any SOTA model after task targeted pre-training. We use attention layers to build a dual alignment, i.e. alignment of meta-data knowledge to queries and strong aligned visual representations. This ensures that the generated answers are personalized and more relevant to the user query. We show significant improvements in 1-hop and spatial performance from previous SOTA models (aligned to our task objective). Our results also note a $\approx$6% increase in accuracies with our training methodology for the chosen architecture, on user's personal data. With an aim to build a system that can truly empower user experience, we also conduct a user study to qualitatively evaluate our model. Our participants found the personalized results significantly more relevant and noted a very high need for integration of proposed solution into the device ecosystems. In future work, we want to extend this model to be trained with a larger data-set (combining other open knowledge data-sets post domain validation), to improve system performance. Our model currently lacks OCR capabilities. We want to build and integrate such a solution to allow the model to "read & answer". Our final target is to build contextual understanding into our model, allowing for a more natural, dialog-like, interaction system.

References

[1] Valeria Fionda and Giuseppe Pirrò. Learning triple embeddings from knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3874–3881, 2020. 3

[2] Diego Garcia-Olano, Yasumasa Onoe, and Joydeep Ghosh. Improving and diagnosing knowledge-based visual question answering via entity enhanced knowledge injection. In *Companion Proceedings of the Web Conference 2022*, pages 705–715, 2022. 2, 6

[3] François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. Conceptbert: Concept-aware representation for visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 489–498, 2020. 2

[4] 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. corr abs/1612.00837 (2016). *arXiv preprint arXiv:1612.00837*, 2016. 4, 7

[5] R Gradilla. Multi-task cascaded convolutional networks (mtcnn) for face detection and facial landmark alignment. *link]. Acessado em*, 13, 2020. 5

[6] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022. 2

[7] Natthawut Kertkeidkachorn and Ryutaro Ichise. T2kg: An end-to-end system for creating knowledge graph from unstructured text. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 3

[8] Guohao Li, Xin Wang, and Wenwu Zhu. Boosting visual question answering with context-aware knowledge aggregation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1227–1235, 2020. 2

[9] Ramprasaath Selvaraju Akhilesh Gotmare Shafiq Joty Caiming Xiong Li, Junnan and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems 34 (2021): 9694-9705.*, 2021. 2

[10] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 2

[11] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 4

[12] Ishani Mondal, Yufang Hou, and Charles Jochim. End-to-end nlp knowledge graph construction. *arXiv preprint arXiv:2106.01167*, 2021. 3

[13] Abhishek Narayanan, Abijna Rao, Abhishek Prasad, and S Natarajan. Vqa as a factoid question answering problem: A novel approach for knowledge-aware and explainable visual question answering. *Image and Vision Computing*, 116:104328, 2021. 5

[14] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33, 2015. 3

[15] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 6

[16] Edward W Schneider. Course modularization applied: The interface system and its implications for sequence control and data analysis. 1973. 3

[17] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, pages 146–162. Springer, 2022. 4

[18] Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8876–8884, 2019. 3, 4

[19] Violetta Shevchenko, Damien Teney, Anthony Dick, and Anton van den Hengel. Reasoning over vision and language: Exploring the benefits of supplemental knowledge. *arXiv preprint arXiv:2101.06013*, 2021. 2

[20] Ajeet Kumar Singh, Anand Mishra, Shashank Shekhar, and Anirban Chakraborty. From strings to things: Knowledge-enabled vqa model that can read and reason. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4602–4612, 2019. 2

[21] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 2

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[23] Peter Vickers, Nikolaos Aletras, Emilio Monti, and Loïc Barrault. In factuality: Efficient integration of relevant facts for visual question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 468–475, 2021. 6, 7

[24] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052*, 2022. 2, 5

[25] Ziqing Yang, Yiming Cui, and Zhigang Chen. Textpruner: A model pruning toolkit for pre-trained language models. *arXiv preprint arXiv:2203.15996*, 2022. 6

[26] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study

of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3081–3089, 2022. 2

[27] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2

[28] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 2

[29] Maryam Ziaeefard and Freddy Lecue. Towards knowledge-augmented visual question answering. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1863–1873, 2020. 5