SkillCLIP: Skill Aware Modality Fusion Visual Question Answering (Student Abstract)

Atharva Naik[*], **Yash Parag Butala**[*], **Navaneethan Vaikunthan**[*], **Raghav Kapoor**

Carnegie Mellon University
{arnaik, ypb, nvaikunt, raghavka}@andrew.cmu.edu

Abstract

When humans are posed with a difficult problem, they often approach it by identifying key skills, honing them, and finally effectively combining them. We propose a novel method and apply it for the VizWiz VQA task to predict the visual skills needed to answer a question, and leverage expert modules to produce intermediary outputs and fuse them in a skill-aware manner. Unlike prior works in visual question-answering (VQA) that use intermediate outputs such as detected objects and Optical Character Recognition (OCR), our approach explicitly guides the model with a skill embedding on what to focus on. While our results show that using skill-aware fusion outperforms skill-unaware models for only a subset of questions, we believe our results provide interesting directions for future work. We also release our code, model, and illustrative demonstrations for future research purposes.

Introduction

VQA has gained traction since the publication of VQAv2 (Goyal et al. 2016). However, progress on the accessibility-focused VizWiz VQA dataset (Gurari et al. 2018) has proven particularly difficult (Cao et al. 2022). The dataset was collected by asking day-to-day questions to visually impaired individuals. Hence, visual evidence in the provided image is often blurry or out of frame and the questions are highly conversational in nature. Their distribution differs from most other VQA datasets affects the transferability of architectures (Cao et al. 2022) pre-trained on datasets like VQAv2. Despite these difficulties, improvements have been made in VQA using multimodal large language models (LLMs) with well over 1B parameters that require extensive pretraining on private datasets. Without such resources, a possible way of tackling this problem is to break the question down into simpler subtasks like a human. However, identifying the sub-tasks and integrating the resultant outputs is nontrivial. Hence, we propose a novel approach to integrating the output modalities of several expert modules by predicting the visual skills needed to correctly answer a given question. We leverage skill annotations in the VizWiz Vision Skills task: Object Recognition, Text Recognition, Color Recognition, and Counting. Instead of explicitly choosing experts, our approach fuses their intermediate outputs together with explicit guidance from learned skill embeddings. Our expert modules for Object, Text, and Color Recognition are a pre-trained object detector, an Optical Character Recognition (OCR) pipeline, and a CLIP Image encoder respectively. We show in our results that our proposed architecture, ***SkillCLIP***, outperforms comparable models using explicit skill guidance for answering.

Methodology

We introduce ***SkillCLIP***, a skill-aware, late modality fusion-based approach that uses CLIP to encode text and image type modalities, as shown in figure 1. We also propose ***SkillCLIP*-multitask**, a joint learning-based approach for learning skill and answer classification simultaneously as shown in 2. For all the experiments, we leverage the CLIP-large model (ViT-L/14) which has image encoding and language encoding of 768 dimensions each.

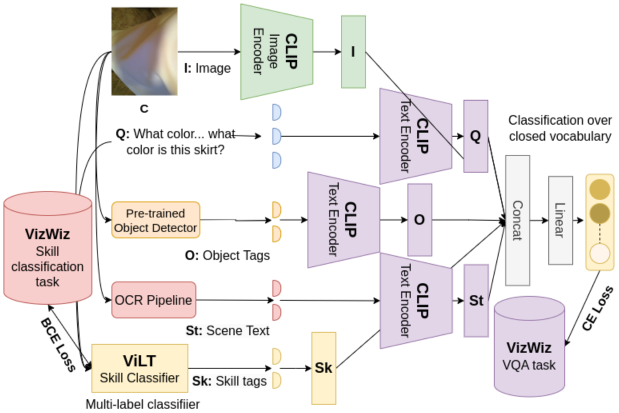

Figure 1: ***SkillCLIP*** Model Architecture. We predict skill embedding using ViLT and use it to fuse different encodings.

SkillCLIP **Architecture** Modalitites are independently encoded by passing them through CLIP and then used for late fusion. It is helpful to keep the modalities separate since CLIP model can encode texts up to 77 tokens. Passing each of the textual inputs separately avoids truncation and loss of data. Description of each module is as follows:

- **Skill prediction:** VizWiz has a skill prediction task[1] where given an image and a question, the task is to predict the visual skills required for answering the question. The train

[*] These authors contributed equally.

[1] https://vizwiz.org/tasks-and-datasets/vision-skills/

Methods	Dev Accuracy ↑	Text Accuracy ↑	Object Accuracy ↑	Color Accuracy ↑	Count Accuracy ↑	Answer BERTSim ($\lambda = 0.7$)
CLIP-Large fusion (Radford et al. 2021)	61.37%	32.8%	39.78%	53.29%	45.05%	64.621%
ViLT (Kim, Son, and Kim 2021)	61.82%	36.05%	45.04%	**61.44%**	41.44%	67.122%
GIT-Large (Wang et al. 2022)	51.24%	32.40%	42.66%	58.45%	26.13%	60.130%
SkillCLIP	62.17%	39.16%	46.58%	57.23%	**50.45%**	68.627%
SkillCLIP Multitask	62.16%	37.64%	44.62%	56.01%	49.55%	68.025%
SkillCLIP (w/o skill embedding)	62.48%	38.38%	46.07%	57.23%	45.95%	68.743%
SkillCLIP (w/o object tags)	**62.58%**	**41.00%**	46.74%	54.58%	45.95%	**69.553%**
SkillCLIP (w/o scene text)	62.31%	40.44%	**47.24%**	56.28%	48.65%	69.160%

Table 1: The table shows accuracy of different models over the dev set as well as skill-wise performance. The last column shows BERTScore similarity of the answers, for relaxed evaluation of the generative model.

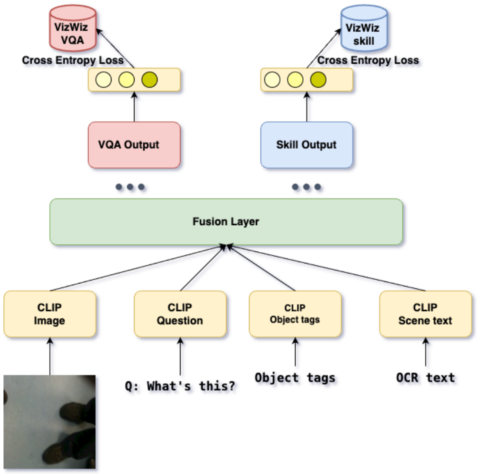

Figure 2: ***SkillCLIP*-Multitask** Model Architecture

sample has 14,259 images. We train ViLT as a skill predictor to label remaining unlabelled images.

- **Scene text:** We use Easy-OCR[2] to collect the scene text in an image and use edit distance to correct extraction errors.
- **Object detection:** We use the DETA model[3] trained on the COCO dataset for object detection.
- **Image encoder:** We obtain the CLIP encoding of the image and pass it to the fusion layer.
- **Question encoder:** The question is encoded through CLIP's language encoder.

Modalities are fused through linear layers.

***SkillCLIP*-Multitask Architecture** We propose a model where skill can be learned as a joint task along with VQA. We motivate a model that can be trained on skill prediction and attend to required modalities accordingly. For skill prediction, we consider the train examples in the VizWiz skill prediction task. We reuse the SkillCLIP expert modules and jointly train on the 20523 VQA questions and 14240 skill prediction instances.

Experiments and Result

We present our results in Table 1. The first section includes the results from finetuning baselines while the next section shows our models and their ablations. GIT-large (Wang et al. 2022) is generative while the rest are classification-based. The results show that ***SkillCLIP*** outperforms the baselines for overall performance. On a skill-level basis, it shows significant improvement of around $\sim 5\%$, 2%, and 5% for questions requiring text, object, and count skills respectively. We also analyze the effect of different expert modules.

Conclusion and Future Work

We present ***SkillCLIP***, a novel framework for VQA, inspired by a human problem-solving approach. ***SkillCLIP*** predicts the necessary visual skills for VQA and leverages expert modules to generate skill-aware intermediary outputs that enhance overall and skill-specific task performance. We also release our code and models for reproduction[4]. In future, we plan to integrate self-attention-based fusion for better integration of modalities. Other directions include better scene-text fusion, finer skill embeddings, and reducing bias in VQA, ultimately enhancing the task for the visually challenged.

References

Cao, Y. T.; Seelman, K.; Lee, K.; and Daumé III, H. 2022. What's Different between Visual Question Answering for Machine" Understanding" Versus for Accessibility? *arXiv preprint arXiv:2210.14966*.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2016. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. *International Journal of Computer Vision*.

Gurari, D.; Li, Q.; Stangl, A. J.; Guo, A.; Lin, C.; Grauman, K.; Luo, J.; and Bigham, J. P. 2018. VizWiz Grand Challenge: Answering Visual Questions from Blind People. *CoRR*, abs/1802.08218.

Kim, W.; Son, B.; and Kim, I. 2021. ViLT: Vision-Language Transformer Without Convolution or Region Supervision. In *International Conference on Machine Learning*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Wang, J.; Yang, Z.; Hu, X.; Li, L.; Lin, K.; Gan, Z.; Liu, Z.; Liu, C.; and Wang, L. 2022. GIT: A Generative Image-to-text Transformer for Vision and Language. *ArXiv*, abs/2205.14100.

[2] https://pypi.org/project/easyocr/

[3] https://github.com/jozhang97/DETA

[4] https://zenodo.org/records/8342662