NorViVQA: Visual Question Answering for Visually Impaired in Norwegian Language

Ratnabali Pal[1,*], Samarjit Kar[1] and Dilip K. Prasad[2]

[1]Department of Applied Mathematics, NIT Durgapur, India
[2]UiT The Arctic University of Norway, Tromsø, Norway

Abstract

Designing a visual question answering (VQA) system for low-resource languages is challenging. Yet, it has enormous potential for practical applications toward making AI-driven assistive technologies more inclusive and accessible. In this article, we introduce Norsk VQA for visually impaired (NorViVQA), a Norwegian VQA dataset derived from the VizWiz-VQA, which contains real-world, often low-quality images captured by blind users alongside questions. Using the NorT5-base model, fine-tuned for English–Norwegian translation, we translate questions and answers into Norwegian Bokmål while preserving the original challenges of VizWiz. We propose a light-weight VQA module on top of the Contrastive Language-Image Pre-training (CLIP) for answer type prediction and answer prediction. We demonstrated the effect of different Norsk embedding methods and achieved 67.09% answer type prediction accuracy and 34.86% answer accuracy. This research aims to bridge the research gap in VQA for visually impaired users in the Norwegian language. The low accuracy of NorViVQA opened up new challenges for the research community. The code and the datasets will be available in http://www.github.com

Keywords

VQA, CLIP-VQA, Visually impaired, low-resource language, NorViVQA

1. Introduction

Dr. Cecily Morrison, a Microsoft researcher who is visually impaired (VI), said, “AI can empower blind and low-vision users by providing real-time, context-aware information about their surroundings." Visually impaired people rely on image and audio based assistive devices to help interpret their surroundings. They need technology that makes everyday tasks, such as identifying products or objects, navigating spaces, reading texts on labels, and recognising scenes, etc., much more accessible. They require assistive technology as their "eyes," converting visual data into accessible formats. Visual Question Answering (VQA) [1], a multi-modal reasoning task in artificial intelligence that enables users to learn about visual content using natural language queries. According to Ethnologue [2], 7,139 languages are officially recognised worldwide, yet most AI breakthroughs, notably VQA, are mostly focused on high-resource languages such as English. This leaves a significant research gap for resource-constrained languages, including Norwegian, where annotated datasets, pre-trained models, and linguistic resources are limited. As a result, accessing AI-driven assistive devices is challenging for many non-English-speaking populations, especially those with disabilities. Although multi-lingual visual question answering (M-VQA) [3] datasets have recently advanced, most of these efforts still fail to adequately represent low-resource languages (those with inadequate computing and linguistic resources). Accessibility may be revolutionized with a multilingual VQA system that includes the majority of the spoken languages for people with disabilities. It can help users become more independent, empower them, and encourage social inclusion by addressing their language, visual, and assistive needs. In this research, we propose the development of a resource-constrained VQA model designed for the Norwegian language within the broader context of multilingual VQA. We emphasise the significance of focusing on these languages,

NAIS 2025: Symposium of the Norwegian AI Society, June 17–18, 2025, Tromsø, Norway
*Corresponding author.
pal.ratnabali@gmail.com (R. Pal); samarjit.kar@maths.nitdgp.ac.in (S. Kar); dilip.prasad@uit.no (D. K. Prasad)
https://github.com/Ratnabali-Pal/ (R. Pal)
1234-5678-9012 (R. Pal); 0000-0002-5503-9338 (S. Kar); 0000-0002-3693-6973 (D. K. Prasad)

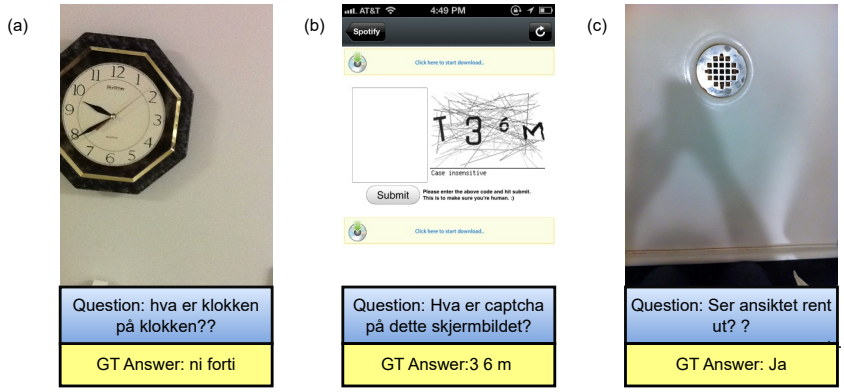

Figure 1: Example image-question-answer pairs from the VizWiz-VQA dataset translated into NorViVQA.

exploring how linguistic diversity, data availability, and AI model adaptability can shape the future of inclusive AI-powered assistive technology.

Traditional Large visual models (LVMs) [4], Large Language Models (LLMs) [5] such as GPT-4 [6], RAG [7], LLaVA [8] are trained on large volumes of text data to generate human-like answers. While impressive in their ability to produce coherent text, these models have a significant drawback: (a) They require large and wide neural networks with millions or billions of parameters to capture complex text or visual data patterns. (b) Models like GPT-4 and LLaVA, for example, need layers of transformers and attention processes, each of which needs a large number of parameters to work well. Here, we address the dual difficulties of low-resource language challenges and computational sustainability to make them a practical option for worldwide deployment. The Norwegian query and visual input are encoded into embeddings using the text and image encoders of Contrastive Language–Image Pre-training (CLIP) [9] model. This model has demonstrated great performance in learning with cross-modal supervision from numerous image-text pairs collected online.

Our research aims to bridge the gap in low-resource languages by developing VQA-based smart solutions that specifically address the needs of visually impaired individuals, allowing greater accessibility and independence in their everyday lives. To address the challenges, we contribute as:

- Using NorT5-base [10], fine-tuned for bidirectional English ↔ Norwegian (Bokmål and Nynorsk) translation, we propose a Norwegian VQA dataset (NorViVQA) generated for low-vision individuals, authenticated from the VizWiz-VQA [11] dataset. This dataset is derived from the VizWiz-VQA corpus, where visually impaired people captured real-world images and asked visual questions based on them. Due to the nature of the data, many of the images are unfocused or poorly framed, reflecting the challenges faced by visually impaired users in capturing meaningful visual content. Figure 1 (a),(b), and (c) show three examples of the Vizwiz-VQA dataset.

- We propose an image-text CLIP pipeline to learn visual concepts with natural language supervision. To evaluate our dataset, we implement a small trainable baseline model that integrates a Vision Transformer (ViT) with NorBERT as the language encoder, trained, and achieved a state-of-the-art accuracy on the NorViVQA dataset.

Such a VQA system can successfully address the world's linguistic diversity, empower visually impaired people, and offer equitable access to information. Our approach facilitates more effective interaction between low-resource language speakers and their surroundings by bridging the language gap in AI-driven assistive devices. As a result, the AI ecosystem becomes more inclusive. It also promotes increased independence, social inclusion, and digital accessibility.

The subsequent sections of this paper are structured as follows: An analysis of existing work on visual question answering in a wide range of applications is provided in Section 2. Section 3 introduces and details the construction of our Norwegian dataset, NorViVQA, derived from the VizWiz-VQA dataset

and translated using the NorT5-base model. Section 4 outlines our proposed approach for Norwegian VQA. Our experimental findings for Norwegian VQA are detailed in Section 5. The concluding remark of our study is provided in Section 6.

2. Related Works

As mentioned in the Introduction, we summarize recent articles focusing on VQA for visually impaired individuals and multilingual VQA applications, including low-resource language line Norsk.

2.1. VQA for VI people

Visual Question Answering (VQA) for Visually Impaired (VI) People focuses on designing systems that help visually impaired individuals gain insights about their surroundings or understand visual content through a combination of image analysis and a question-answering pipeline. Individuals with visual impairments may be unable to check the content of their captured images, leading to reduced image quality. As far as we know, VizWiz [11]dataset is the largest VQA dataset specially designed for low-vision people. VizWiz has more than 31,000 visual questions created by low-vision people who each capture an image using a mobile phone and record a spoken question about it. Each visual question has ten crowdsourced answers. In this study [12], researchers created a new VQA system that combines implicit textual knowledge and implicit multimodal knowledge using encoder-decoder generative models. In this article, the VizWiz-VQA-Grounding dataset [13] presented a base model for designing less biased VQA models and more accurate answer grounding models, making it a significant milestone for future research.

2.2. Low-resource and Multilingual VQA

Here, we summarise visual question answering for additional languages, including low- and high-resource languages, along with models and resources. Since many visually impaired people may not speak English well, it is preferable to use a VQA system in their mother tongue to promote better communication and understanding.

Chinese: Wang et al. presented a novel bilingual scene (text + evidence) VQA data set named EST-VQA [14] which is annotated with English and Chinese QA pairs. In this article, three challenges (cross-language, localisation, and tradition) are proposed to assess the generalisation of VQA models.

Vietnamese: Researchers have presented a new dataset for Vietnamese VQA called ViVQA [15], which includes 15,000 question-answer pairs in Vietnamese and 10,328 images to assess Vietnamese VQA models.

Hindi: Hindi is the official language of the Indian government (along with English) and is widely spoken in North and Central India. A few years back, researchers designed a unified end-to-end framework for multilingual and code-mixed question answering and introduced a dataset [16] for Hindi and code-mixed VQA.

Multilingual BERT: BERT (Bidirectional Encoder Representations Transformers) [17] is a foundational model in the field of NLP that introduced a novel approach to pre-training language models. BERT processes text in both directions (left-to-right and right-to-left) in each layer, allowing it to construct more context-aware representations of words. Additionally, BERT is trained to predict whether one sentence follows another, helping it understand the relationships between sentences and making it effective for tasks requiring contextual sentence-level understanding. Devlin et al. proposed Multilingual BERT (M-BERT) [17], which follows the same architecture and training procedure as BERT, but it is pre-trained as a single language model on the concatenation of monolingual Wikipedia corpora from 104 languages. After a large number of probing experiments, researchers have concluded that while M-BERT [18] does learn multilingual representations, these representations display systematic deficiencies, particularly affecting certain language pairs.

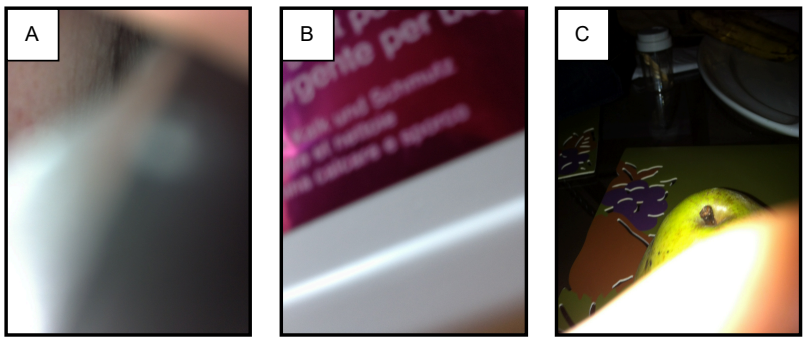

Figure 2: Examples of challenging images in the VizWiz dataset. (A) Noisy image resulting from shaky camera movement; (B) Blurred image caused by camera focusing issues; (C) Image with partially captured object, making it difficult to interpret.

2.3. Norwegian QA

While most Norwegian datasets focus on text-based Question Answering (QA), they lack the multimodal components needed for VQA, limiting their applicability in image-based AI tasks. Developed for machine reading comprehension, NorQuAD [19] contains over 4,752 manually curated QA pairs in Bokmål and Nynorsk. NorOpenBookQA [20], NorCommonSenseQA [20], NorTruthfulQA [20], and NRK-Quiz-QA [20] datasets are designed for different reasoning tasks, including common sense and factual knowledge. Pretrained Norwegian language models like NorBERT [10, 21] and NorT5 [10] have been fine-tuned for QA tasks. It is noted that there are no such VQA datasets in the Norwegian language.

3. Proposed Dataset

The literature study on VQA for visually impaired people suggests that the VizWiz dataset is one of the best datasets for the task. VizWiz is a pioneering VQA dataset introduced by Gurari et al. [11], designed specifically to reflect real-world accessibility challenges. It contains over 31,000 visual questions collected from visually impaired individuals, who captured photos using mobile phones and asked questions about them. Each visual question is paired with 10 crowdsourced answers to ensure response diversity. The dataset is also equipped with four classes (answer type), namely “others", “numbers", “yes/no", and “unanswerable". The VizWiz dataset has two characteristics that are different from traditional VQA datasets as:

- The images are often of low quality, as they are captured by low-vision individuals using mobile phones, leading to frequent issues such as poor framing, blurriness, or inadequate lighting.
- Due to either poor image quality or a lack of relevant visual information, a significant proportion of the visual questions are inherently unanswerable.

All of these features collectively make VizWiz a uniquely challenging and practical benchmark for developing VQA systems for accessibility. Building upon this, we propose NorViVQA, a Norwegian VQA dataset designed for low-vision individuals. In this paper, we utilize the ltg/nort5-base-en-no-translation model—an encoder-decoder architecture [22] based on the T5 framework [23]—specifically optimized for machine translation [24] between English (en) and the one official written standard of Norwegian: Bokmål (nb). This model enables the creation of a custom English-to-Norwegian translation pipeline, allowing us to convert existing English-language datasets, such as VizWiz-VQA, into Norwegian. By leveraging this translation capability, we generate a Norwegian version of the dataset suitable for VQA applications targeted toward low-vision individuals. The distributions of the questions for English and Norwegian are presented in Figure 3.

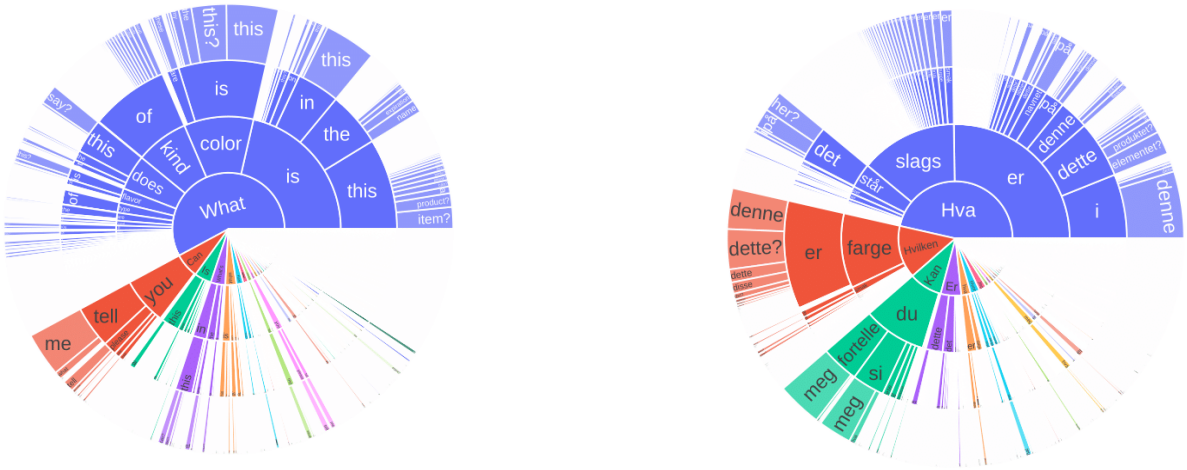

Figure 3: Sunburst diagram showing the distribution of the first four words in questions from the VizWiz-VQA (English on the left) and NorViVQA (Norwegian on the right) datasets.

4. Proposed Method

Here, we propose to use a small trainable neural network on top of a pre-trained image and language encoder for the VQA. VQA aims to answer a textual question based on the content of an image by using an image's content. We use CLIP [25, 26]architecture as a baseline. CLIP can function as a base by:

- The image and question are encoded into the same embedding space.
- Predicting a suitable answer using a baseline neural network considering the joint embeddings.

Let us consider the NorViVQA dataset $D = \{I_i, Q_i\}_{i=1}^{N}$. The goal is to predict an answer, A. The image is encoded into a vector embedding $z_I \in \mathbb{R}^d$. Both the Questions (Q_K) and user answers A_K are encoded into the same language embedding space, $z_{Q_K} = g[(Q)_K]$ and $z_{A_K} = g[(A_K)]$ for $K = 1, 2, ..., k$.

The image and Norwegian questions/answers are embedded into the CLIP model. CLIP ensures that matching Norwegian image-text pairs are aligned into a shared embedding space. Here, our proposed model consists of three main modules: (a) a pre-trained image encoder, (b) a pre-trained text encoder, and (c) a baseline neural network for answer prediction. The proposed method is depicted in Figure 4. These modules are discussed hereafter.

4.1. Image and Text Encoder

We propose to use an image encoder (ViTL/14) [27, 28]from OpenAI to extract visual features. The Vision Transformer is trained with contrastive pretraining on image-text pairs. This encoder transforms input images into a dense embedding that captures semantic information aligned with natural language, making it suitable for image-text alignment tasks. We applied it to the NorViVQA dataset to obtain semantically rich image embeddings that align well with text, supporting tasks involving visual question answering for visually impaired users. ViT-L/14 is the large variant of the Vision Transformer (ViT) model, where the input image is divided into non-overlapping patches of size 14×14 pixels. Each patch is flattened and linearly projected into a fixed-size embedding vector, typically of 1024 dimensions for the large (L) variant. To retain spatial information, learnable positional embeddings are added to the patch embeddings. The model architecture consists of 24 Transformer encoder blocks, each comprising multi-head self-attention [29], Layer Normalization [30], and a feedforward neural network with GELU activation [31]. Additionally, a learnable classification token ([CLS]) is prepended to the sequence of patch embeddings. The final representation of this token is used as the global image embedding. In the context of CLIP, the output image embeddings from ViT-L/14 are aligned with text embeddings

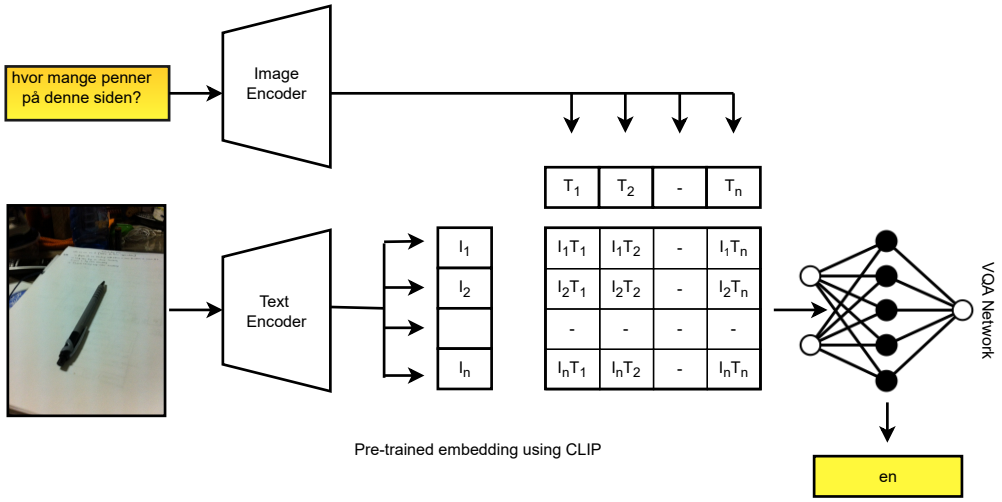

Figure 4: Overview of our proposed CLIP pipeline that learns visual concepts from natural language supervision by jointly training an image encoder and a text encoder to align image-text pairs in a shared embedding space.

in a shared latent space through contrastive learning, enabling strong vision-language understanding across diverse tasks.

In our research, ltg/norbert3-large [32, 10, 21] is used to generate text embeddings that are rich in contextual and linguistic cues of the Norwegian language, allowing effective alignment with visual features in multimodal tasks. ltg/norbert3-large is a BERT-based language model developed specifically for the Norwegian language as part of the NorBERT3 series. The ltg/norbert3-large model follows the BERT-large architecture, consisting of 24 Transformer layers, 16 attention heads, and a hidden size of 1024. It is pre-trained using the Masked Language Modeling (MLM) [33] objective on extensive Norwegian corpora, including news, books, web content, and parliamentary texts. The model uses a WordPiece tokeniser [34] designed for Norwegian vocabulary and generates contextualized token embeddings, with the [CLS] token commonly employed for sentence-level tasks such as classification or similarity scoring.

4.2. Baseline Model

A text encoder (NoRBERT) [10, 21] and an image encoder (ViT) [35] are combined in a single space and used as the input in the proposed baseline VQA module.

The baseline consists of layer normalization, dropout, and linear layers. The answer type is predicted using a linear layer, and the answers are predicted using a Sigmoid activation layer, as shown in Figure 5. As we stated earlier, the image and text embeddings are combined into a single space, although they are embedded differently. The layer normalization helps to normalize across features; it stabilized the training process and also reduced the dependency of the batch size. The dropout here is 0.5, it will reduce the overfitting and also increase the generalization across different training data. It is noted that only this module is trained here.

5. Results and discussion

Following the discussion of our proposed approach, we will analyze our experimental results.

We have reported the results of varying text embeddings on the same baseline trainable module.

Experimental Setup: All experiments were performed on an Intel Xeon Silver server equipped with 128 GB of RAM and a 24GB RTXA5000 GPU. The dataset consists of 20523 training samples that are split into 80% training and 20% validation sets. The test data consists of 1243 samples. Training was performed using the Early Stopping criterion with a patience of 15. We employed a batch size of 64 and

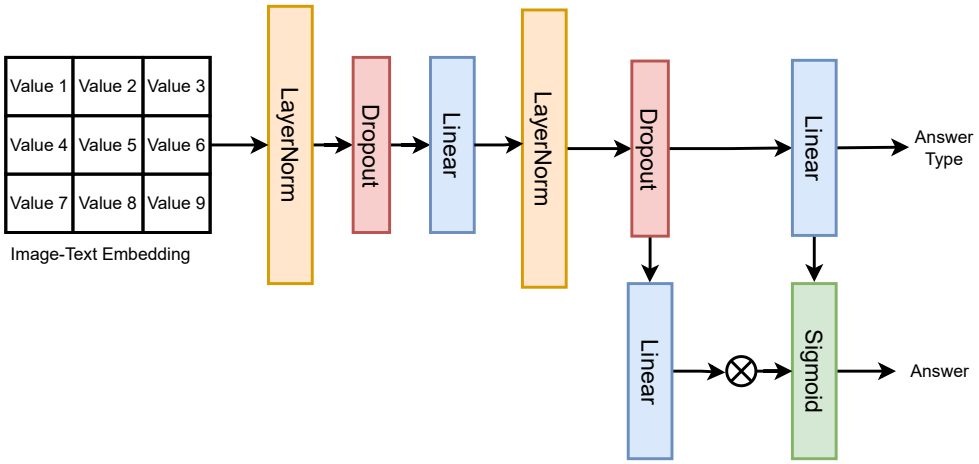

Figure 5: Architecture of the proposed VQA module.

a learning rate 1×10^{-4} with the Adaptive Momentum Optimizer. State-of-the-art approaches were implemented utilizing the PyTorch and Keras libraries. Training is stopped if the model's performance on the validation set fails to improve for 10 consecutive epochs, ensuring stopping before overfitting. The dataset has two tasks: (a) TYPE accuracy, i.e., identification of four answer types, and (b) ANSWER prediction accuracy. The loss is calculated as:

$$L = \frac{1}{2}\left(A_{type} + A_{answer}\right)$$

Where: - L is the loss. - A_{type} is the answer type accuracy. - A_{answer} is the answer accuracy.

Results of NorViVQA: Table 1 presents a comparative evaluation of four models—ltg/norbert3-large, NorGLM/NorGPT-3B, OpenAI-GPT, and M-CLIP/XLM-Roberta-Large-Vit-L-14— VQA task on the NorViVQA dataset. The ltg/norbert3-large model achieves the best performance across all metrics, with a test TYPE accuracy of 67.09%, Test ANS accuracy of 34.86%, and an average accuracy of 50.97%, indicating strong generalization and answer accuracy. It also yields the lowest Test Loss (120.97), showing more stable training. OpenAI-GPT performs moderately in predicting answer types (29.02%) but completely fails in answer correctness (0.00% ANS accuracy), resulting in a lower average accuracy (14.51%) and higher test loss (170.73). M-CLIP/XLM-Roberta-Large-ViT-L-14 records the lowest performance, with a Test TYPE Accuracy of 15.68%, ANS Accuracy of only 0.0006%, and an Average Accuracy of 7.84%. Its test loss is highest at 176.81, indicating difficulty in adapting to the VQA for VI task. These results demonstrate the effectiveness of domain-specific fine-tuning (as seen in ltg/norbert3-large) and highlight the limitations of general-purpose large vision-language models (like OpenAI-GPT) when directly applied to specialised tasks like video-based VQA for visually impaired users, particularly when text extraction and contextual understanding are crucial.

Failure Cases and Model Improvement: It is noted that the accuracy of both answer type prediction and answer prediction has a scope of improvement. The failure cases can be from incorrect image embedding, as the pre-trained CLIP model is trained using English image-text pairs (MS-COCO). A complete fine-tuning of the CLIP model or the ViT-based image encoder on the Norwegian language can improve the accuracy. On the other hand, the Norwegian language embedding can also be fine-tuned using large volumes and diverse texts.

5.1. Ablation Study

We have tested with two other VQA network variations to understand the proposed baseline's effectiveness. One variation consists of 34 layers similar to the baseline and achieved 67.1% test type accuracy, which is only 0.01% better than the baseline. The test answer accuracy did not change. Another variation

Table 1
Results of proposed baseline VQA module varying different text embeddings

Text Embedding	Metric	Score
ltg/norbert3-large	Test TYPE Accuracy (%)	67.09
	Test ANS Accuracy (%)	34.86
	Average TEST Accuracy (%)	50.97
	Test Loss	120.97
NorGLM/NorGPT-3B	Test TYPE Accuracy (%)	67.09
	Test ANS Accuracy (%)	34.50
	Average TEST Accuracy (%)	51.12
	Test Loss	122.25
Openai-gpt	Test TYPE Accuracy (%)	29.02
	Test ANS Accuracy (%)	0.06
	Average TEST Accuracy (%)	14.51
	Test Loss	170.73
M-CLIP/XLM-Roberta-Large-Vit-L-14	Test TYPE Accuracy (%)	15.68
	Test ANS Accuracy (%)	0.00
	Average TEST Accuracy (%)	7.84
	Test Loss	176.81

is that we use multi-head attention [36] with the baseline. We achieved 67.11% test type accuracy, which is 0.02% better, and 34.89% test answer accuracy, which is 0.03% better than the baseline. Changing the batch size to 8 and 16 did not change the performance.

6. Conclusion

In this article, we propose a new dataset called Norwagiean VQA for visually impaired (NorViVQA). First, we translate the English questions from the VizWiz dataset into Norwegian using a pre-trained ltg/nort5-base-en-no-translation model. Next, the image and text embedding are fed into the CLIP model for further processing. Finally, we employ a small trainable VQA module for question-answer. Our proposed approach highlights the potential of combining language translation and CLIP models in a small VQA module and addresses inclusivity for Norwegian-speaking users. Our work enhances support for low-vision or visually impaired people by enhancing contextual comprehension in visual question-answering, allowing them to interact with visual content in their local language more successfully.

In the future, we hope to expand this technology to support multilingual question processing, making it applicable to a wider range of languages and user groups. This extension will increase the accessibility and usability of VQA solutions for visually impaired people.

Declaration on Generative AI

During the preparation of this work, the author(s) used Grammerly in order to: Grammar and spelling check. The author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

Acknowledgement

Funding for this work was given by the HORIZON-ERC-POC Project (Spermotile, Project ID: 10112725), the H2020 Project (OrganVision, Project ID: 964800), HORIZON Europe (BETTER, Project ID: 101136262)

and the Research Council of Norway Project (nanoAI, Project ID: 325. (With reference to resources and hardware).

References

[1] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, M. Rohrbach, Towards vqa models that can read, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 8317–8326.

[2] H. Hammarström, Ethnologue 16/17/18th editions: A comprehensive review, Language 91 (2015) 723–737.

[3] Y. Wang, J. Pfeiffer, N. Carion, Y. LeCun, A. Kamath, Adapting grounded visual question answering models to low resource languages, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2596–2605.

[4] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, J. Gao, Vinvl: Revisiting visual representations in vision-language models, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 5579–5588.

[5] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, et al., Chatgpt for good? on opportunities and challenges of large language models for education, Learning and individual differences 103 (2023) 102274.

[6] R. Mao, G. Chen, X. Zhang, F. Guerin, E. Cambria, Gpteval: A survey on assessments of chatgpt and gpt-4, arXiv preprint arXiv:2308.12488 (2023).

[7] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in Neural Information Processing Systems 33 (2020) 9459–9474.

[8] H. Liu, C. Li, Y. Li, Y. J. Lee, Improved baselines with visual instruction tuning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 26296–26306.

[9] S. Shen, L. H. Li, H. Tan, M. Bansal, A. Rohrbach, K.-W. Chang, Z. Yao, K. Keutzer, How much can clip benefit vision-and-language tasks?, arXiv preprint arXiv:2107.06383 (2021).

[10] D. Samuel, A. Kutuzov, S. Touileb, E. Velldal, L. Øvrelid, E. Rønningstad, E. Sigdel, A. Palatkina, Norbench–a benchmark for norwegian language models, NEALT Proceedings Series (2023) 618–633.

[11] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, J. P. Bigham, Vizwiz grand challenge: Answering visual questions from blind people, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3608–3617.

[12] Y. Sun, Q. Si, Z. Lin, W. Wang, T. Mei, Outside-knowledge visual question answering for visual impaired people, in: 2023 IEEE International Conference on Medical Artificial Intelligence (MedAI), IEEE, 2023, pp. 49–54.

[13] C. Chen, S. Anjum, D. Gurari, Grounding answers for visual questions asked by visually impaired people, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 19098–19107.

[14] X. Wang, Y. Liu, C. Shen, C. C. Ng, C. Luo, L. Jin, C. S. Chan, A. v. d. Hengel, L. Wang, On the general value of evidence, and bilingual scene-text visual question answering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10126–10135.

[15] K. Q. Tran, A. T. Nguyen, A. T.-H. Le, K. Van Nguyen, Vivqa: Vietnamese visual question answering, in: Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation, 2021, pp. 683–691.

[16] D. Gupta, P. Lenka, A. Ekbal, P. Bhattacharyya, A unified framework for multilingual and code-mixed visual question answering, in: Proceedings of the 1st conference of the Asia-Pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing, 2020, pp. 900–913.

[17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers

for language understanding, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 4171–4186.
[18] T. Pires, E. Schlinger, D. Garrette, How multilingual is multilingual bert?, arXiv preprint arXiv:1906.01502 (2019).
[19] S. Ivanova, F. A. Andreassen, M. Jentoft, S. Wold, L. Øvrelid, Norquad: Norwegian question answering dataset, arXiv preprint arXiv:2305.01957 (2023).
[20] V. Mikhailov, P. Mæhlum, V. O. C. Langø, E. Velldal, L. Øvrelid, A collection of question answering datasets for norwegian, arXiv preprint arXiv:2501.11128 (2025).
[21] A. Kutuzov, J. Barnes, E. Velldal, L. Øvrelid, S. Oepen, Large-scale contextualised language modelling for norwegian, in: Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa), 2021, pp. 30–40.
[22] T. M. Doan, D. Baumgartner, B. Kille, J. A. Gulla, Automatically detecting political viewpoints in norwegian text, in: International Symposium on Intelligent Data Analysis, Springer, 2024, pp. 242–253.
[23] K. Grover, K. Kaur, K. Tiwari, Rupali, P. Kumar, Deep learning based question generation using t5 transformer, in: Advanced Computing: 10th International Conference, IACC 2020, Panaji, Goa, India, December 5–6, 2020, Revised Selected Papers, Part I 10, Springer, 2021, pp. 243–255.
[24] H. Wang, H. Wu, Z. He, L. Huang, K. W. Church, Progress in machine translation, Engineering 18 (2022) 143–153.
[25] L. Fan, D. Krishnan, P. Isola, D. Katabi, Y. Tian, Improving clip training with language rewrites, Advances in Neural Information Processing Systems 36 (2023) 35544–35575.
[26] M. Hafner, M. Katsantoni, T. Köster, J. Marks, J. Mukherjee, D. Staiger, J. Ule, M. Zavolan, Clip and complementary methods, Nature Reviews Methods Primers 1 (2021) 20.
[27] J. Chen, Q. Yu, X. Shen, A. Yuille, L.-C. Chen, Vitamin: Designing scalable vision models in the vision-language era, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 12954–12966.
[28] K. Habel, F. Deuser, N. Oswald, Clip-reident: Contrastive training for player re-identification, in: Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports, 2022, pp. 129–135.
[29] C. Wu, F. Wu, S. Ge, T. Qi, Y. Huang, X. Xie, Neural news recommendation with multi-head self-attention, in: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), 2019, pp. 6389–6394.
[30] J. Xu, X. Sun, Z. Zhang, G. Zhao, J. Lin, Understanding and improving layer normalization, Advances in neural information processing systems 32 (2019).
[31] M. Lee, Mathematical analysis and performance evaluation of the gelu activation function in deep learning, Journal of Mathematics 2023 (2023) 4229924.
[32] E. Rønningstad, L. C. Storset, P. Mæhlum, L. Øvrelid, E. Velldal, Mixed feelings: Cross-domain sentiment classification of patient feedback, arXiv preprint arXiv:2501.19134 (2025).
[33] N. Arefyev, D. Kharchev, A. Shelmanov, Nb-mlm: Efficient domain adaptation of masked language models for sentiment analysis, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 9114–9124.
[34] D. Schönle, C. Reich, D. O. Abdeslam, Linguistic-aware wordpiece tokenization: Semantic enrichment and oov mitigation, in: 2024 6th International Conference on Natural Language Processing (ICNLP), IEEE, 2024, pp. 134–142.
[35] F. Bao, S. Nie, K. Xue, Y. Cao, C. Li, H. Su, J. Zhu, All are worth words: A vit backbone for diffusion models, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 22669–22679.
[36] J.-B. Cordonnier, A. Loukas, M. Jaggi, Multi-head attention: Collaborate instead of concatenate, arXiv preprint arXiv:2006.16362 (2020).