A Deep Learning Approach for Medical Visual Question Answering via Cascaded Multitask Learning

Teodora-Alexandra Toader, Alexandru Manole, Gabriela Czibula

WeADL 2025 Workshop

The workshop is organized under the umbrella of WinDMiL, project funded by CCCDI-UEFISCDI, project number PN-IV-P7-7.1-PED-2024-0121, within PNCDI IV

Contents

Introduction and Motivation

- Multimodal Vision-Language Learning fuses visual and textual information to enable models with more robust reasoning and real-world applicability.

- Visual Question Answering (VQA), and its medical counterpart MVQA, leverage this fusion to interpret clinical images and answer medical questions in natural language. MVQA can assist healthcare professionals by providing clinical decision support and improving access to medical information.

- Challenges of MVQA: limited amount of data that can make it easier for models to overfit and not provide enough generalization

Background - MVQA

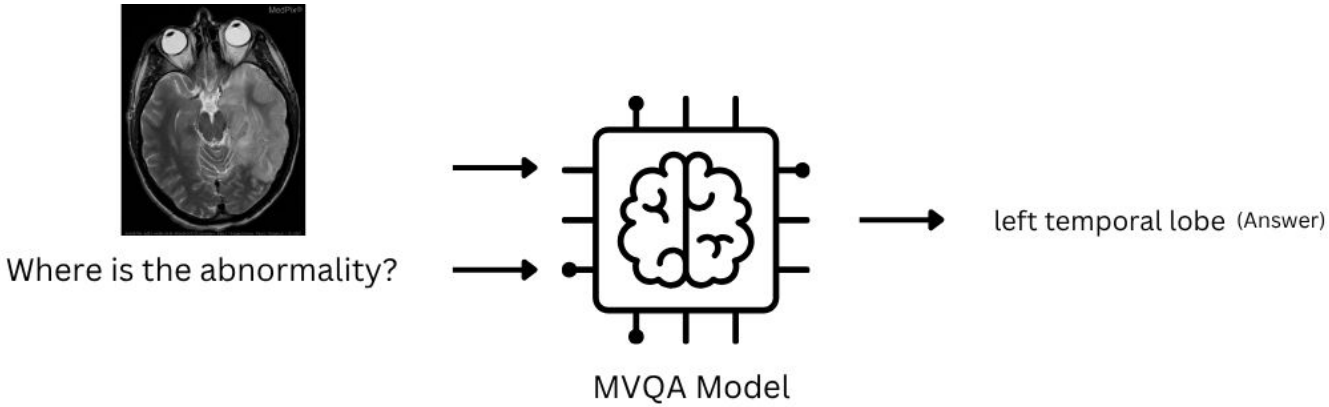

Figure 1: Overview of the MVQA Task.

Background - Datasets

- **OVQA Dataset [HWLH22]**
  - dataset that focuses on orthopedic medical data
  - 2001 images and 19 020 QA pairs
  - Questions are divided into six different categories: abnormality, attribute other, conditions presence, modality, organ system and plane
- **VQA-Med-2019 dataset [AH+19]**
  - 4200 images selected from MedPix database and 15 292 corresponding questions divided into train, validation and test data
  - Questions were divided into four different categories: Organ, Plane, Modality and Abnormality
  - Both closed-ended and open-ended types of questions

Background - Related Work

MVQA Tasks: Tackled as both classification and generation problems.

Architecture Trends: Earlier methods use separate encoders and fusion (e.g., MLP, attention) [Y+19, VS+19], while more recent work adopts transformer-based models [RZ20, KB+21, CDW+23].

Question-Type Splitting: Some models divide tasks by question type (e.g., organ, plane, abnormality) [ZKR19, AT+19], some using classifiers for factual queries and generators for open-ended ones.

Multitask & Meta-Learning: Techniques like Skeleton-based Sentence Mapping and MEVF modules with MAML and CDAE address data limitations [L+20, N+19, DN+21].

Generative & LLM-Based Models: Models such as [VSD+23] generate open-form answers using visual prompts and LLMs (e.g., GPT-2).

Background - Multitask Learning

- **What is Multitask Learning?**
  - Multi-task Learning (MTL) is a popular paradigm in which models are trained to perform multiple tasks simultaneously by sharing representations.

- **Challenges of MTL**
  - Choosing the right tasks
  - Designing the appropriate architecture
  - Determining the optimal way to combine losses during learning

Background - MTL in VQA/MVQA

VQA Reframed as MTL: Pollard et al. [PS20] train a multitask model across question types (e.g., object, numeric, colour, spatial), outperforming single-task baselines.

SFN Model for MVQA: Kornuta et al. [KRS+19] introduce Supporting Facts Network with five classification heads, combining fused input and auxiliary knowledge (e.g., image size).

MTL in Pretraining: Gong et al. [GCL+21] pretrain an image encoder via semantic tasks (e.g., segmentation) and an image-question compatibility task to enhance performance on VQA-RAD [LG+18].

Caption-Based MTL Framework: Cong et al. [CXGT22] add a captioning task and question-type classification heads for both image and text encoders, improving multimodal understanding.

Approach

- We propose a multitask model named **CAMMA** for the MVQA task that uses other annotations that most MVQA datasets have, such as question type, answer type, and organ type in a cascaded multitask architecture to reduce overfitting and improve generalization.

Approach - Research Questions

- Does multitask learning work as a method to improve generalization and reduce overfitting for models developed for solving MVQA?

- Does the additional information embedded in our multitask approach lead to an enhanced performance of the model?

- Would symbiotic tasks for an MVQA multitask approach be useful for increasing model accuracy compared to the single-task approach?

Approach - Task Formulation

- Denoting by *I* a set of medical images, by *Q* a set of questions/texts in natural language, and by *C* a set of classes corresponding to the considered tasks (in our case the *main answer*, the *image organ type*, the *answer type*, and the *question type*), the MVQA task in a multi-task formulation can be formalized as the problem of learning a mapping: $\Phi : I \times Q \rightarrow C$

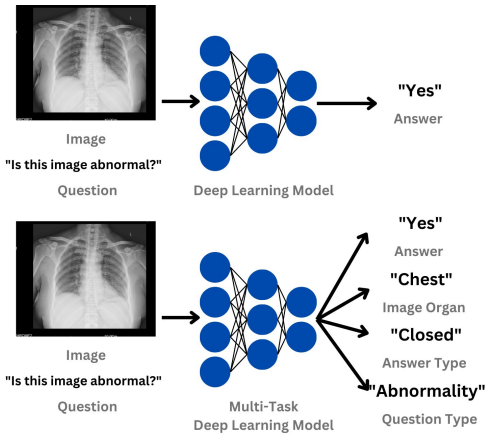

Figure 2: Overview of the MVQA task in the normal and multitask formulation of the task.

Approach - General Architecture

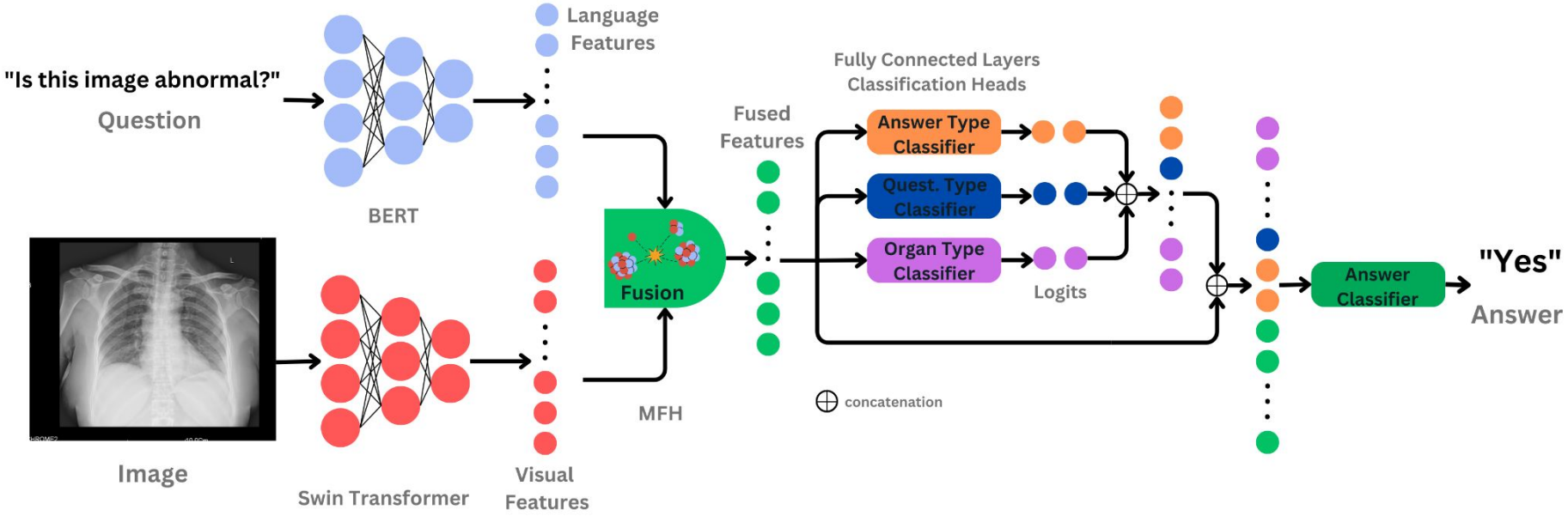

Figure 3: Overview of the proposed CAMMA model om the OVQA dataset.

Approach - General Architecture

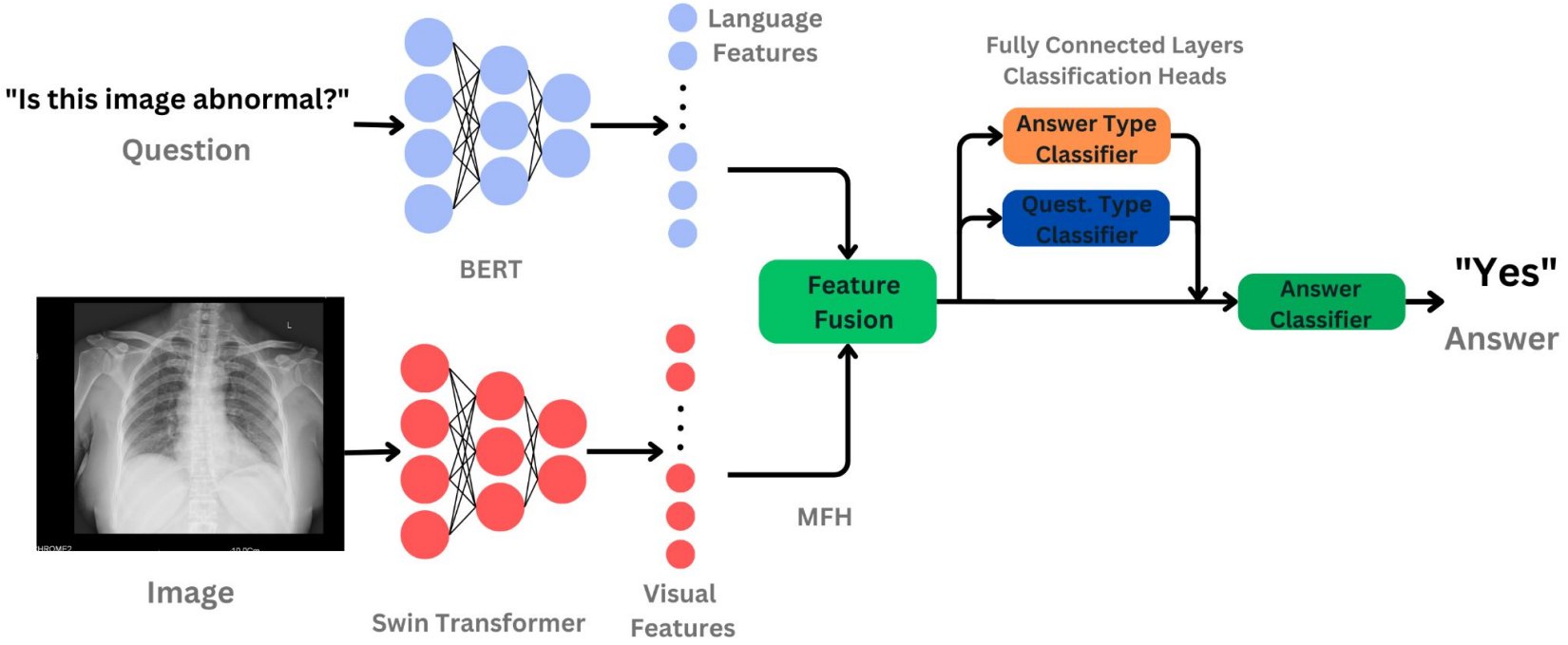

Figure 4: Overview of the proposed CAMMA model om the VQA-Med 2019 dataset.

Experimental Results - Architecture

Image Encoder	Text Encoder	Fusion Strategy	Multitask	OVQA	
				Accuracy	BLEU
SWIN	BERT	MFH	X	0.6230	0.6784
			✓	**0.7145**	**0.7559**
		MFB	X	0.5846	0.6398
			✓	0.674	0.7194
VGG19		MFH	X	0.5962	0.6543
			✓	0.6451	0.6979
		MFB	X	0.5588	0.6143
			✓	0.6161	0.6731
ViT		MFH	X	0.6119	0.6683
			✓	0.6803	0.7305
		MFB	X	0.5799	0.636
			✓	0.6424	0.7012

Table 1: Performance of CAMMA based on the choice of image encoder, fusion module and use of MTL.

Experimental Results - Task selection

main answer	question type	answer type	image organ	OVQA Test accuracy
✓				0.623
✓	✓			0.6524
✓		✓		0.6335
✓			✓	0.6482
✓	✓	✓		0.6766
✓	✓		✓	0.6992
✓		✓	✓	0.6824
✓	✓	✓	✓	**0.7145**

Table 2: Results on the OVQA dataset different classification tasks selection.

main answer	question type	answer type	VQA-Med-2019 Test Accuracy
✓			0.552
✓	✓		0.558
✓		✓	**0.568**
✓	✓	✓	0.562

Table 3: Results on the VQA-Med 2019 dataset different classification tasks selection.

Experimental Results - Comparison to Related Work

Model	Accuracy
Our CAMMA model	**0.7145**
MEVF-SAN [N+19]	0.6190
MEVF-BAN [N+19]	0.6100
MMQ-SAN [DN+21]	0.6850
MMQ-BAN [DN+21]	0.650
MMBERT [KB+21]	0.6330
PTUnifier [CDW+23]	0.7130
Generative LLM [VSD+23]	0.7100
OpenAI's GPT-4o	0.3123

Table 4: Comparison to related work on OVQA dataset.

Experimental Results

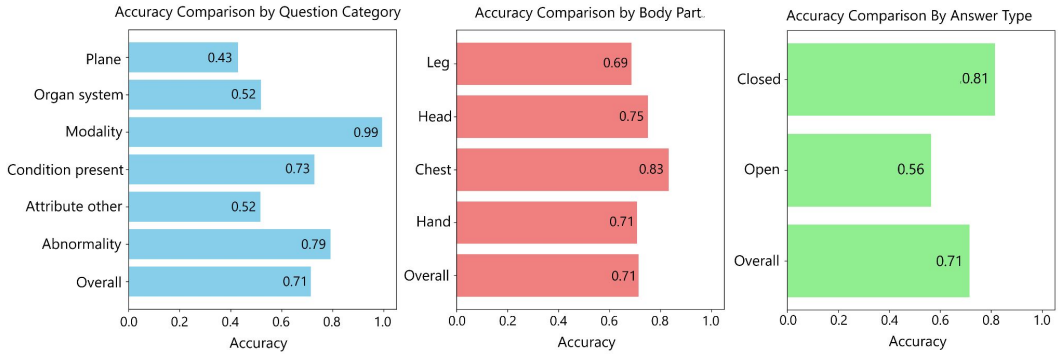

Figure 5: Performance comparison based on question category, body part, and answer type on OVQA Dataset.

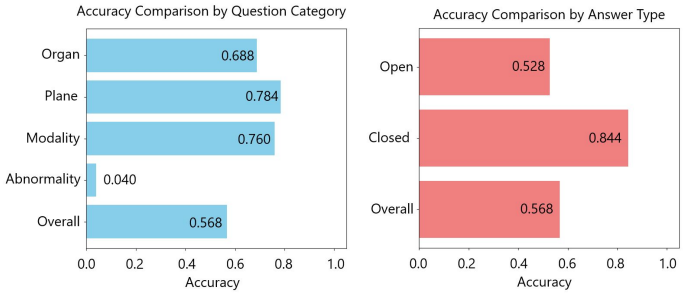

Figure 6: Performance comparison based on question category and answer type on VQA-Med-2019.

Conclusions and Future Work

We presented CAMMA, a cascading multitask architecture for Medical Visual Question Answering that achieved state-of-the-art results on the OVQA dataset, showing that multitask learning improves performance and reduces overfitting.

Our experiments confirm that embedding additional information through multiple classification heads helps address data scarcity, and a cascaded approach consistently increases performance by leveraging knowledge of organ, question and answer types.

For future work, we plan to use task weights as hyperparameters to optimize task balance, explore new tasks to add to the framework, and investigate self-supervised tasks such as image or question reconstruction due to limited extra annotations. We also plan to look into explainability and interpretability for future models.

References

[AH+19] Asma Ben Abacha, Sadid A. Hasan, et al. VQA-Med: Overview of the Medical Visual Question Answering Task at ImageCLEF 2019. In *Working Notes of CLEF 2019*, volume 2380 of *CEUR Workshop Proceedings*, 2019.

[AT+19] Aisha Al-Sadi, Bashar Talafha, et al. JUST at ImageCLEF 2019 Visual Question Answering in the Medical Domain. In *Working Notes of CLEF 2019*, volume 2380, 2019.

[CDW+23] Zhihong Chen, Shizhe Diao, Benyou Wang, Guanbin Li, and Xiang Wan. Towards unifying medical vision-and-language pre-training via soft prompts. In *Proceedings of ICCV'23*, pages 23403–23413, 2023.

[CXGT22] Fuze Cong, Shibiao Xu, Li Guo, and Yinbing Tian. Caption-aware medical VQA via semantic focusing and progressive cross-modality comprehension. In *Proceedings of MM'22*, pages 3569–3577, 2022.

[DN+21] Tuong Do, Binh X Nguyen, et al. Multiple meta-model quantifying for medical visual question answering. In *Proceedings of MICCAI 2021: Part V 24*, pages 64–74. Springer, 2021.

[GCL+21] Haifan Gong, Guanqi Chen, Sishuo Liu, Yizhou Yu, and Guanbin Li. Cross-modal self-attention with multi-task pre-training for medical visual question answering. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, pages 456–460, 2021.

[HWLH22] Yefan Huang, Xiaoli Wang, Feiyan Liu, and Guofeng Huang. OVQA: A clinically generated visual question answering dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2924–2938, 2022.

[KB+21] Yash Khare, Viraj Bagal, et al. MMBERT: Multimodal BERT pretraining for improved medical VQA. In *Proceedings of ISBI 2021*, pages 1033–1036. IEEE, 2021.

[KRS+19] Tomasz Kornuta, Deepta Rajan, Chaitanya Shivade, Alexis Asseman, and Ahmet S Ozcan. Leveraging medical visual question answering with supporting facts. arXiv preprint arXiv:1905.12008, 2019.

[L+20] Zhibin Liao, et al. AIML at VQA-Med 2020: Knowledge inference via a skeleton-based sentence mapping approach for medical domain visual question answering. In *CLEF (Working Notes)*, pages 1–14, 2020.

References

[LG+18] Jason J Lau, Soumya Gayen, et al. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5(1):1–10, 2018.

[N+19] Binh D Nguyen, et al. Overcoming data limitation in medical visual question answering. In *Proceedings of MICCAI 2019, Part IV 22*, pages 522–530, 2019.

[PS20] Amelia Elizabeth Pollard and Jonathan L Shapiro. Visual question answering as a multi-task problem. arXiv preprint arXiv:2007.01780, 2020.

[RZ20] Fuji Ren and Yangyang Zhou. CGMVQA: A New Classification and Generative Model for Medical Visual Question Answering. *IEEE Access*, 8:50626–50636, 2020.

[SLR19] Lei Shi, Feifan Liu, and Max P Rosen. Deep Multimodal Learning for Medical Visual Question Answering. In *CLEF (Working Notes)*, 2019.

[VS+19] Minh Vu, Raphael Sznitman, et al. Ensemble of Streamlined Bilinear VisualQA Models for the ImageCLEF 2019 Challenge in the Medical Domain. In *CLEF 2019*, volume 2380, pages 1–11, 2019.

[VSD+23] Tom Van Sonsbeek, Mohammad Mahdi Derakhshani, et al. Open-ended medical visual question answering through prefix tuning of language models. In *Proceddings of MICCAI 2023*, pages 726–736, 2023.

[Y+19] Xin Yan, et al. Zhejiang University at ImageCLEF 2019. In *Working Notes of CLEF 2019*, volume 2380, pages 1–9, 2019.

[ZKR19] Yangyang Zhou, Xin Kang, and Fuji Ren. TUA1 at ImageCLEF 2019 VQA-Med: a Classification and Generation Model based on Transfer Learning. In *CLEF (Working Notes)*, 2019.

Thank you!