Medical Image Analysis

Volume 99, January 2025, 103357

Article preview

https://doi.org/10.1016/j.media.2024.103357

A Foundation Language-Image Model of the Retina (FLAIR): encoding expert knowledge in text supervision

Julio Silva-Rodríguez [a], Hadi Chakor [b], Riadh Kobbi [b], Jose Dolz [a c], Ismail Ben Ayed [a c]

Highlights

- FLAIR: A vision-language foundation model for fundus images with an assembly dataset.

- Encoding expert’s knowledge in text descriptions.
- Excellent properties for zero-shot generalization.
- Domain-specific foundation models outperform larger-scale generalists models.
- FLAIR model weights and adaptation are made publicly available.

Abstract

Foundation vision-language models are currently transforming computer vision, and are on the rise in medical imaging fueled by their very promising generalization capabilities. However, the initial attempts to transfer this new paradigm to medical imaging have shown less impressive performances than those observed in other domains, due to the significant domain shift and the complex, expert domain knowledge inherent to medical-imaging tasks. Motivated by the need for domain-expert foundation models, we present FLAIR, a pre-trained vision-language model for universal retinal fundus image understanding. To this end, we compiled 38 open-access, mostly categorical fundus imaging datasets from various sources, with up to 101 different target conditions and 288,307 images. We integrate the expert’s domain knowledge in the form of descriptive textual prompts, during both pre-training and zero-shot inference, enhancing the less-informative categorical supervision of the data. Such a textual expert’s knowledge, which we compiled from the relevant clinical literature and community standards, describes the fine-grained features of the pathologies as well as the hierarchies and dependencies between them. We report comprehensive evaluations, which illustrate the benefit of integrating expert knowledge and the strong generalization capabilities of FLAIR under difficult scenarios with domain shifts or unseen categories. When adapted with a lightweight linear probe, FLAIR outperforms fully-trained, dataset-focused models, more so in the few-shot regimes. Interestingly, FLAIR outperforms by a wide margin larger-scale generalist image-language models and retina domain-specific self-supervised networks, which emphasizes the potential of embedding experts’ domain knowledge and the limitations of generalist models in medical imaging. The pre-trained model is available at: https://github.com/jusiro/FLAIR.

Introduction

At least 1 billion people have a vision impairment that could have been prevented or is yet to be addressed (WHO, 2019). In this context, color fundus images combined with computer vision systems present a promising, cost-effective solution for population-based screenings and early detection of ophthalmologic diseases (Balyen and Peto, 2019, Bellemo et al., 2019).

Driven by public datasets, deep learning has reached remarkable performances in a breadth of fundus image analysis problems, such as diabetic retinopathy grading (Liu et al., 2022), glaucoma detection (Orlando et al., 2019), lesion segmentation (Porwal et al., 2020) or multi-disease detection (Cen et al., 2021). Nevertheless, several limitations impede the widespread adoption of these methods. In particular, current deep learning solutions for fundus image analysis may not generalize well whenever there are shifts in the imaging data or in the task at hand (e.g., new or rare classes) (Li et al., 2021, Sengupta et al., 2020). In retinal imaging, and in the much broader field of medical imaging, the current dominant deep learning paradigm is to supervise models on very specific tasks, e.g., diabetic retinopathy classification into a few grades (Liu et al., 2022). Learning representations that might be too specialized for the task and training images at hand, such task-focused models may have difficulty in (i) dealing with the high variability in real clinical scenarios (Finlayson et al., 2021), due to the high variations in image acquisition and patient demographics; and (ii) capturing rare conditions that are not well represented in the training data.

There is currently a paradigm shift in artificial intelligence algorithms, driven by the growing prevalence of models trained on large and diverse datasets, which could be adapted to a broad span of downstream tasks. These models, commonly referred to as *foundation* models, have gained increasing popularity and showed significant success in computer vision and natural language processing tasks (Brown et al., 2020, Radford et al., 2021). In particular, vision-language models such as CLIP (Radford et al., 2021) or ALIGN (Jia et al., 2021) have shown impressive generalization capabilities when fine-tuned on various downstream computer-vision tasks, emerging as powerful alternatives to narrowly-supervised, task-focused models. Learning from large-scale amounts of image–text pairs, such models leverage the rich semantic knowledge in the language-based supervision, thereby yielding visual features that are more descriptive than their task-specific counterparts.
In computer vision tasks, such as image classification, this new *pretrain-and-finetune*paradigm enhanced robustness to image-data shifts and showed promising zero-shot and few-shot transferability. Nonetheless, initial attempts to directly apply these foundation models to the medical domain yielded less convincing performances (Wang et al., 2022c). Indeed, generalist models like CLIP may not capture the fine-grained image features and class dependencies/hierarchies, which might be complex, highly specialized concepts inherent to the expert's domain knowledge; see Fig. 1 for an illustration in the case of retinal fundus images. This has recently motivated the development of foundation models specialized for medical imaging applications (Wójcik, 2022, Moor et al., 2023).
Vision-language models are currently emerging in medical image analysis. Several recent studies investigated foundation models specialized in radiology (Zhang et al., 2022b, Huang et al., 2021b, Wang et al., 2022c), focusing mostly on chest-radiography data. These were motivated by the prevalence of diagnostic text reports in radiology, and the availability of large domain resources to mine such textual information (Bodenreider, 2004, Jain et al., 2021). However, this may not be the case in other medical-imaging modalities. In retinal imaging, for instance, text information is scarce and most datasets are categorically labeled (see Table 1), *i.e.*, the label of each training image is a single category (or class), e.g., "*mild diabetic retinopathy*" (mildDR).
We argue that, even for categorically-labeled images, vision- language pre-training is an appealing solution to integrate domain-specific, fine-grained knowledge, such as the dependencies between the categories, into visual representations. The analysis of medical images by clinical experts is a process

of searching for differential features of candidate conditions. In this process, there are, for instance, hierarchical dependencies between the presence of local lesions and the differential diagnosis at the global level. Such expert's domain knowledge is usually overlooked in conventional training but could be integrated in the form of text descriptions, to build powerful image-language models. To illustrate this, we provide in Fig. 2 a few retinal-imaging examples with categorical labels along with the corresponding text descriptions encoding domain knowledge. For instance, the text description "*only a few microaneurysms are present*" informs on local conditions known to point to the category mildDR (Wilkinson et al., 2003). In Table S4, we provide a comprehensive list of the correspondences between the categorical labels and textual domain-knowledge descriptions, which we compiled from the relevant clinical literature and community standards (Wilkinson et al., 2003, Garner and Ashton, 1979, Allen and Vasavada, 2006, Gass, 1988, Hamel, 2006, Ruiz-Medrano et al., 2019, Yang et al., 2016), to build our foundation model of the retina.

In this work, we introduce FLAIR, a Foundation LAnguage-Image model of the Retina, for color fundus image analysis. FLAIR is trained and validated on a large assembly of datasets, with 288,307 images and different target categories, which we compiled from different publicly available sources. We integrate the expert's domain knowledge in the form of text supervision during both pre-training and zero-shot prediction, thereby enhancing the categorical information of the data. Such a textual expert's knowledge describes the fine-grained features of the pathologies as well as the hierarchies and dependencies between them. We report comprehensive evaluations, comparisons and ablation studies, which show the substantial effect of embedding expert knowledge and the strong generalization and transferability capabilities of FLAIR under challenging scenarios with domain shifts or novel (unseen) categories. When adapted with a lightweight linear probe classifier, FLAIR outperforms models that are fully trained on the target dataset, more so under low-data (few-shot) settings. Furthermore, FLAIR outperforms by a large margin more generalist, larger-scale image-language models such as CLIP or BiomedCLIP. Our results point to the potential of embedding expert domain knowledge and to the limitations of generalist models.

Section snippets

Transfer learning in medical image analysis

Training robust deep-learning models from scratch requires large datasets and huge computational resources (Erhan et al., 2009). These conditions are rarely met in medical imaging. The high variability in image acquisition, the low prevalence of certain conditions, and the limited resources of institutions make it difficult for standard supervised-learning models to capture the substantial variability in real clinical contexts. This is due to the fact that supervised models are typically

Methodology

Fig. 3 depicts an overview of our framework. We introduce each methodological component formally in the following sections.

Assembling the dataset .

A total of 38 public datasets are assembled for training and evaluating the proposed universal model. A summary of the datasets is presented in Table 1. The assembled dataset combines the main tasks explored for fundus image analysis, which include: diabetic retinopathy grading (Decencière et al., 2014, Porwal et al., 2020, Castillo Benítez et al., 2021, Takahashi et al., 2017, Lin et al., 2020, Li et al., 2019b, Nakayama et al., 2023), glaucoma detection (Li et al., 2019a, Kovalyk et al., 2022

Generalization

In this section, we evaluate the generalization capabilities of the proposed model by direct prediction (*i.e.* no adaptation of the trainable parameters) on several target datasets under two common scenarios: *(i)* domain shift, where the set of classes remains the same, but images present domain drifts, and *(ii)* novel classes, where the domain remains the same, but unseen classes are expected to be identified.

Discussion

We introduced FLAIR, a novel vision-language foundation model for universal pathology detection and classification in retinal fundus images. Encoding expert's domain knowledge in the form of text-prompt supervision, FLAIR is trained on an assembly of publicly available, mostly categorical datasets, containing up to different target categories. By leveraging domain knowledge, we mitigated the scarcity of text-based supervision in retinal fundus imaging datasets, opening a promising avenue

CRediT authorship contribution statement

Julio Silva-Rodríguez: Conceptualization, Data curation, Investigation, Methodology, Validation, Visualization, Writing – original draft. **Hadi Chakor:** Data curation, Supervision, Validation. **Riadh Kobbi:** Funding acquisition, Project administration, Validation. **Jose Dolz:** Conceptualization, Formal analysis, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Ismail Ben Ayed:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Project

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The work of J. Silva-Rodríguez was partially funded by the *Fonds de recherche du Québec*(FRQ) under the Postdoctoral Merit Scholarship for Foreign Students (PBEEE). The work is supported, in part, by PROMPT Québec , via its PARTNERSHIP-AI program. We also thank Calcul Québec and Compute Canada.

References (116)

- AraújoT. et al.

  DRGRADUATE: Uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images

  Med. Image Anal.

  (2020)
- BalyenL. et al.

  Promising artificial intelligence–machine learning–deep learning algorithms in ophthalmology

  Asia-Pac. J. Ophthalmol.

  (2019)
- BellemoV. et al.

  Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study

  Lancet Digit. Health

  (2019)
- CarmonaE.J. et al.

  Identification of the optic nerve head with genetic algorithms

  Artif. Intell. Med.

  (2008)
- Castillo BenítezV.E. et al.

  Dataset from fundus images for the study of diabetic retinopathy

  Data Brief

  (2021)
- DecencièreE. et al.

  TeleOphta: Machine learning and image processing methods for teleophthalmology

  IRBM

  (2013)
- FarnellD.J. et al.

  Enhancement of blood vessels in digital fundus photographs via the application of multiscale line operators

  J. Franklin Inst.

  (2008)
- GiancardoL. et al.

Exudate-based diabetic macular edema detection in fundus images using publicly available datasets
Med. Image Anal.
(2012)
- HassanT. et al.
Deep structure tensor graph search framework for automated extraction and characterization of retinal layers and fluid pathology in retinal SD-OCT scans
Comput. Biol. Med.
(2019)
- KrauseJ. et al.
Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy
Ophthalmology
(2018)