Learning Retinal Representations from Multi-modal Imaging via Contrastive Pre-training

Authors
Emese Sükei*, Elisabeth Rumetshofer**,
Niklas Schmidinger**, Ursula Schmidt-Erfurth*,
Günter Klambauer**, Hrvoje Bogunović*

Affiliations
* OPTIMA Lab, Department of Ophthalmology, Medical University of Vienna, AT
** LIT AI Lab, Institute for Machine Learning, Johannes Kepler University, AT

Introduction

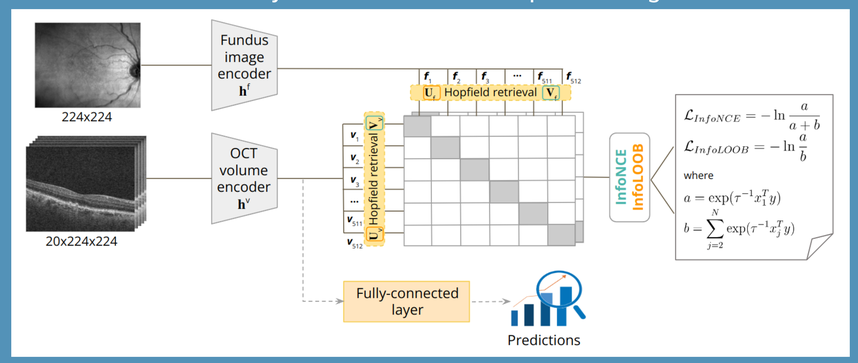
In ophthalmology, large multi-modal datasets are conveniently accessible as retinal imaging scanners acquire 2D fundus images and 3D optical coherence tomography (OCT) for disease evaluation.

Objective

Motivated by this, we propose a multimodal *CLIP* [1] / *CLOOB* [2] objective-based model to learn joint representations of the two retinal imaging modalities, which can then be used for diverse downstream tasks.

Methodology

The study uses large-scale data from OPTIMA Lab imaging datasets for pre-training the encoders. We use the *HARBOR* trial [3] as an external dataset for the downstream tasks. Our framework uses *ResNet18* and *VideoResNet18* with *pre-trained ImageNet and Kinetics* weights as backbone encoders and employs *InfoNCE* and *InfoLOOB objectives* for contrastive pre-training.



For downstream tasks, the volume encoder is used to extract the latent representations, while an additional single fully-connected layer on top is trained to perform the prediction tasks.

Related literature

1. Radford A, et al. Learning Transferable Visual Models From Natural Language Supervision. 2021. Available: http://arxiv.org/abs/2103.00020
2. Fürst A, et al. CLOOB: Modern Hopfield Networks with InfoLOOB Outperform CLIP. 2021. Available: http://arxiv.org/abs/2110.11316
3. Busbee BG, et al. Twelve-month efficacy and safety of 0.5 mg or 2.0 mg ranibizumab in patients with subfoveal neovascular age-related macular degeneration. Ophthalmology. 2013;120: 1046–1056.
4. Kay W, et al. The Kinetics Human Action Video Dataset. 2017. Available: http://arxiv.org/abs/1705.06950

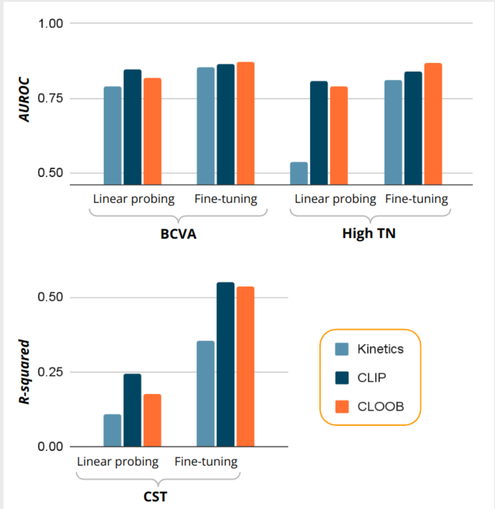
Downstream tasks

We define three downstream tasks on the external dataset, namely predicting:

1. *central subfield thickness* (CST)
2. *best-corrected visual acuity* (BCVA)
3. *high treatment need* (TN).

The first is treated as a regression task, while the latter two as binary classification tasks.

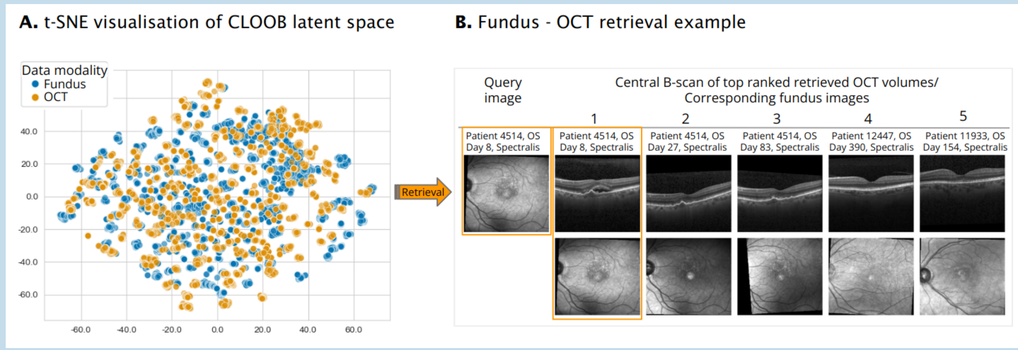
To demonstrate the models' feature extractor capabilities, first, we perform *linear probing*; hence, the encoder weights are kept frozen, and only the prediction layer is trained. Then the models are *fine-tuned* on the same tasks. We use *5-fold cross-validation* to evaluate the models and report the mean scores over the folds.



The obtained mean performance scores on the downstream tasks using the baseline and CLIP/CLOOB pre-trained encoders, respectively.

Retrieval

Top-1 accuracy:
CLIP - 10.51%
CLOOB - 11.36%



Conclusion

Our initial findings suggest that contrastive pre-training with multi-modal retinal images yields *transferable and meaningful OCT volume representations*, which can be leveraged for other clinical tasks.