Unsupervised Representation Learning of Dynamic Retinal Image Changes by Predicting the Follow-up Image

Antoine Rivail* **Hrvoje Bogunović** **Sebastian M. Waldstein**

Bianca S. Gerendas **Wolf-Dieter Vogl** **Ursula Schmidt-Erfurth**

Christian Doppler Laboratory for Ophthalmic Image Analysis
Department of Ophthalmology and Optometry
Medical University of Vienna

1 Introduction

Longitudinal imaging allows to capture both, the *static* anatomical structures and the *dynamic* changes of the morphology due to aging or disease progression. However, common supervised or unsupervised methods for medical imaging do not consider dynamic aspects and process longitudinal data as individual data points. For natural images, algorithms already exist that learn a representation from videos [Walker et al., 2016]. In retinal imaging, however, the temporal sampling resulting from follow-up to disease progression is much lower than in videos. Predictions are therefore more ambiguous and prone to noise

We propose a deep learning approach to overcome these challenges, which allows us to understand the underlying morphological organization and its changes over time, and to discover abnormalities and pathologic evolutions. Our data-driven approach learns a feature representation from unlabeled longitudinal images by predicting the unobserved subsequent image within a series of observations. Several sources of noise, such as imaging noise, misalignment of follow-up images or motion artifacts aggravates the direct prediction of the target image. Thus, we propose to adapt a Conditional Variational Autoencoder (CVAE) [Kingma and Welling] to learn representative static and dynamic features that are robust to noise and uncertainty.

2 Method

Theory of conditional variational autoencoders

Let $X_i = [x_{i-J}, \dots x_i]$ be a sequence of J consecutive images x_i, from which a subsequent unseen image x_{i+1} is predicted. Following [Walker et al., 2016, Kingma and Welling], the optimization of variational autoencoder is based on the variational inequality, which maximizes the likelihood of the prediction $\mathrm{P}(x_{i+1}|X_i)$ by optimizing the last term of the equation 1. Q is the introduced distribution, from which the latent noise vector z is sampled.

$$\begin{aligned} \log \mathrm{P}(x_{i+1}|X_i) &\geq \log \mathrm{P}(x_{i+1}|X_i) - \mathcal{KL}[\mathrm{Q}(z|X_i, x_{i+1})||\,\mathrm{P}(z|X_i, x_{i+1})] \\ &= \mathrm{E}_{z\sim\mathrm{Q}}[\log \mathrm{P}(x_{i+1}|z, X_i)] - \mathcal{KL}[\mathrm{Q}(z|X_i, x_{i+1})|\,\mathrm{P}(z|X_i)] \end{aligned} \tag{1}$$

The term $\mathrm{E}_{z\sim\mathrm{Q}}[\log \mathrm{P}(x_{i+1}|z, X_i)]$ encourages the model to output a correct prediction, where X_i is encoded by the Encoder network. This term is optimized with a L_2 reconstruction loss. The second term, the $\mathcal{KL}$ divergence, forces Q to be a normal distribution ($\mathcal{N}(0, I)$). Thus, the encoding network is forced to extract as much information as possible from the previous images. At testing time, the Q network is discarded, and $z \sim \mathcal{N}(0, I)$.

*Supported by the Austrian Federal Ministry of Science, Research and Economy, and the National Foundation for Research, Technology and Development

1st Conference on Medical Imaging with Deep Learning (MIDL 2018), Amsterdam, The Netherlands.

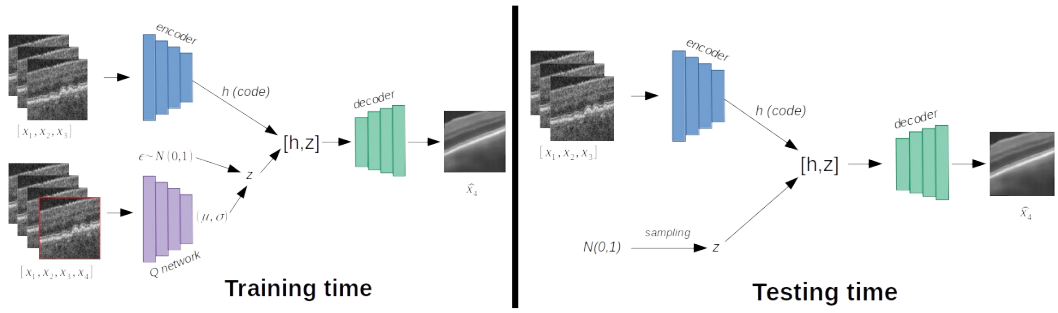

Figure 1: Architecture: the *encoder*, the *Q Network* (μ and σ) and *decoder* are implemented by convolutional networks. At training time $z = \mu + \epsilon \cdot \sigma$, at testing time: $z \sim \mathcal{N}(0, I)$

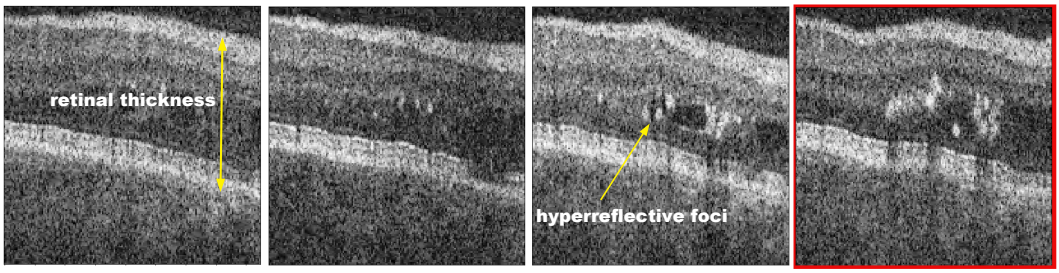

Figure 2: Example of four subsequent OCT B-scans acquired in a monthly interval. The first three images are used as input and the last one as target (red). *Retinal thickness*, *hyperreflective foci* or *Drusens* are morphological properties tested for feature evaluation.

CVAE is computing a distribution over the possible solutions instead of a single prediction as for instance an autoencoder provides. This distribution is built from two parameters: the code vector, $\mathbf{h}$, which encodes deterministic factors and the $\mathbf{z}$ vector, which represents unpredictable factors.

Architecture: the optimization relies on three convolutional networks: (1) the *encoder* network, which extracts important information for predictions in the code vector h, (2) the *Q network* for implementing a standard distribution $N(\mu(X_i, x_{i+1}), \sigma(X_i, x_{i+1}))$, and (3) the *decoder* having as inputs the vectors h and z and which outputs a prediction of the subsequent image (Figure 1).

3 Experiments and results

3.1 Dataset

The dataset contains 3900 OCT scans from 204 different patients diagnosed with intermediate age-related macular degeneration (AMD). Each patient was scanned with a monthly follow-up for a period of up to 24 months. Time-points where a patient already converted to late stage AMD were excluded.

Preprocessing: We built sequences of four consecutive visits without overlapping, and cropped local patches from the central 3 mm of the retina. In order to crop patches from the same anatomical position, we registered all scans to a reference scan within the sequence as described by Vogl et al. [2017]. Finally, every sample is a sequence of four 170 x 170 patches. The first three images were used as inputs to predict the last one (Figure 2).

3.2 Training

We divided the dataset into training and validation subsets for the training of the CVAE and a test set for the final evaluations. Care has been taken that all image series of a patient were in the same subset. The network was trained by stochastic gradient descent using ADAM algorithm. Overfitting was controlled by examining the reconstruction loss on the test set (without Q network).

Table 1: Prediction results of morphological properties. "Direct" shows the results predicted from previous values of the properties only. "CVAE" uses in addition the code vector h from previous images. The target values are scaled to have 0 mean and $unit$ variance (standard scaling).

Method	Retinal thickness	Drusen Volume	HRF volume
R^2 score			
Direct	**0.959**	0.945	0.807
CVAE	0.945	**0.974**	**0.823**
Mean absolute error (standard scaled)			
Direct	**0.152**	0.228	0.406
CVAE	0.177	**0.185**	**0.392**

3.3 Evaluation

In order to evaluate the features produced by the encoder (**h** code), we predicted morphological properties from it, which change over time in intermediate stage of AMD. We used average *total retinal thickness*, *drusen volume* and *hyperreflective foci* (HRF) volume, which were automatically segmented using the methods described in [Garvin et al., 2009, Schlegl et al., 2017].

The prediction target was the average value of a morphological property from the last image in the sequence (unobserved image). We trained a self-normalizing MLP regressor [Klambauer et al., 2017]. For comparison, we performed a regression based on the measured property values from the initial time-points serving as baseline (*Direct*). In the CVAE method, we in addition included the code vector h as regression input. Results are listed in Table 1.

Initial results showed that the prediction code h improves the results of a direct prediction, both for *Drusen Volume* and *HRF* volume, by increasing R^2 score and reducing the mean absolute error. This indicates that the proposed method is able to successfully encode dynamic properties of OCT images. The prediction of *total retinal thickness* was not improved but given that in intermediate AMD total retinal thickness usually remains very stable, it can be predicted more easily by simply regressing from previous thickness measurements.

References

M. K. Garvin, M. D. Abramoff, X. Wu, S. R. Russell, T. L. Burns, and M. Sonka. Automated 3-d intraretinal layer segmentation of macular spectral-domain optical coherence tomography images. *IEEE Transactions on Medical Imaging*, 28(9):1436–1447, Sept 2009.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. URL `http://arxiv.org/abs/1312.6114`.

Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. *CoRR*, abs/1706.02515, 2017. URL `http://arxiv.org/abs/1706.02515`.

Thomas Schlegl, Sebastian M. Waldstein, Hrvoje Bogunovic, Franz Endstraßer, Amir Sadeghipour, Ana-Maria Philip, Dominika Podkowinski, Bianca S. Gerendas, Georg Langs, and Ursula M Schmidt-Erfurth. Fully automated detection and quantification of macular fluid in oct using deep learning. *Ophthalmology*, 2017.

W. D. Vogl, S. M. Waldstein, B. S. Gerendas, U. Schmidt-Erfurth, and G. Langs. Predicting macular edema recurrence from spatio-temporal signatures in optical coherence tomography images. *IEEE Transactions on Medical Imaging*, 36(9):1773–1783, Sept 2017.

Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert. An Uncertain Future: Forecasting from Static Images using Variational Autoencoders. *arXiv:1606.07873 [cs]*, June 2016. URL `http://arxiv.org/abs/1606.07873`.