Prediction of AMD Stage From OCT Scans With 3D Vision Transformers

Sean Cohen
Department of Computer Science and Computational Biology
The Hebrew University of Jerusalem
shon.cohen@mail.huji.ac.il

Abstract

We present a usage of 3D vision transformers in order to predict the features of age-related macular degeneration (AMD) from retinal optical coherence tomography (OCT). We used a partially tagged large real patient's private dataset (1M images). Based on the observation that some features of AMD are irreversible, we largely expand our tagged data. In this paper, we focus on prediction of AMD stage with the Timesformer model developed for video learning tasks such as action recognition [2]. We refer each OCT scan as a video frames. Our model is able to distinguish between dry and wet AMD with accuracy of 84% with AUC and precision of 0.9 and 0.94 respectively. We did not test our model on public datasets yet, but other models achieved accuracy greater than 90% on several datasets for this task. We think that the difference is resulted from many tagging errors and from the fact that our current model takes only low-resolution images. We are truly believe that we can improve the model by keeping explore it. Code and model are available at https://github.com/seanco-hash/AMD_prediction/tree/master

1 Introduction

Age-related macular degeneration (AMD) is a worldwide leading cause of blindness among older adults [1]. AMD is traditionally classified into two types: dry AMD, in which the typical lesions are drusen deposited in the macula, and wet AMD, which is characterized by choroidal neovascularization (CNV). The classification of AMD, particularly the diagnosis of choroidal neovascularization (CNV) activity, is closely associated with its treatment [3]

Optical coherence tomography (OCT) provides clear information on the size, location, and extent of the drusen, as well as the presence and activity of CNV [1]. Several studies have recognized OCT as the most reliable and reproducible tool to evaluate lesion activity, assist clinical doctors in making decisions about treatment, and observe the subsequent therapeutic effect [4]. Retinal OCT scan produces a 3D model of the retina (length X width X height) and allows a 2D view of width strips of the retina layers. A single strip provides a thin view across the retina (width X height). The strips are produced along the retina, giving us a 3D model. In this work, we refer the model as a video, when each strip is a single frame.

Video analysis is a major subject in the field of computer vision. As in the whole field, CNNs are the leading approach of solving a variety of tasks in video analysis in the last years. In the last year, the idea of self-attention mechanism, developed mainly in the NLP field, was applied successfully for vision tasks in images [5]. The applications of these ideas into 3D vision tasks were not late to come, and ones are argue that the self-attention mechanism is natural for video analysis task because of the sequentiality of the data, which is similar to the textual data in NLP [2]. In this work, we use a self-attention mechanism based model called Timesformer [2], which is based on ViT [5]. We hoped that the abilities of Timesformer to analyze successfully related sequential frames would benefit the

Preprint. Under review.

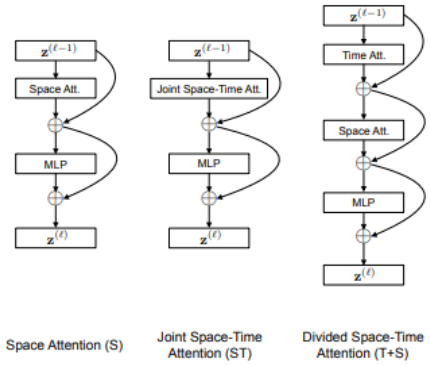

Figure 1: The available video self-attention blocks in Timesformer.

analysis of OCT scans and take into consideration knowledge that lays along the retina and is hidden from human eye when looking at images across the retina.

In practice, the conversion of OCT scans (e2e files) into image sequences causes reduced quality and is less continuous. This limits the self-attention mechanism performance on the temporal space. Despite these limitations, Timesformer is able to focus the self-attention mechanism in the spatial space (means that the most of attention occurs for each frame with itself) and then reach to conclusion based on signals from all of the frames. Although results are not sufficient yet, this work demonstrates that vision transformers have to be considered in analyzing 3D medical data.

2 Preliminary

This section focus on Timesformer [2], which is the model that our work is based on. Developed by Facebook research group, Timesformer is a 3D version of ViT [5]. We chose this model because it was more accessible and fit our schedule limitations. The main problem we have working with Timesformer is the input low resolution (224 X 224) it demands, while our data consists high-resolution images. We assume that using Video Swin [6], which based on Swin Transformer [7], will give us better results because its ability to take high-resolution images as input and because its superiority over Timesformer [6].

2.1 Timesformer

The TimeSformer takes as input a clip X (H×W×3×F) consisting of F RGB frames of size H × W sampled from the original video (OCT scan in our case). Similar to Vit, Timesformer decomposes each frame into N non-overlapping patches, each of size P x P, followed by linear embedding, query-key-value computation and self-attention computation. Then Timesformer encodes together the self-attention outputs resulted from F frames and classify.

2.2 Different types of self-attention blocks and our usage

Timesformer developers explored several mechanisms of self-attention blocks (fig 2.1). Because of the high computational cost of vision transformers, the tried to deal with the trade-off between emphasizing self-attention within each frame (spatial) to self-attention across the different frames (temporal). Finally Timesformer was released with several self-attention blocks option, when 'Divided Space-Time Attention' shows the best results. We were hoping to use the fact that OCT scans provides us continuous images of the retina, which supposed to be convenient input for divided space-time attention model. Unfortunately, the fact that more than 8 or 16 frames input is computationally expensive for Timesformer and after processing the OCT scans into images we have left with not very continuous frames, made the space-time attention model less suitable for us. Instead, we finally used the 'Space Attention' mechanism that focus the attention within each frame. Although inferior from space-time divided attention, this model has shown good results in video analysis tasks too.

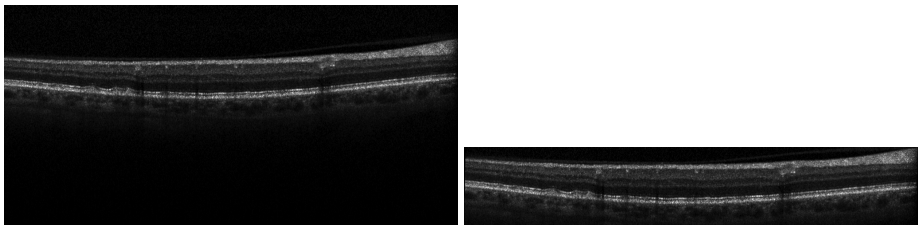

Figure 2: OCT scan single strip before (left) and after (right) preprocessing .

3 Related Work

There are few works trying to automate tasks related to AMD, including few deep learning approaches. Unfortunately, patents usually protect successful approaches and the researchers does not share their exact methods nor code.

3.1 Early AMD detection with CNN

The field of automated detection, classification and segmentation of AMD biomarkers from OCT scans is widely explored [8, 9, 10, 11, 12]. In one prominent work, several computational tools were used together with CNN [13]. The authors made extensive image pre-processing work in order to ease the job and apply inductive bias into the CNN. They used a model based on few versions of ResNet, pretrained on ImageNet. They finetune their model with private dataset consists about 20K OCT Scans but only 10% of them were contain features of disease. Their CNN models were able to identify presence of early AMD pathologies with accuracy ranged from 86%-89%.

3.2 Attention based network for AMD diagnoses

As in the whole computer vision field, approaches to improve or replace the CNN models with attention-based models were developed for AMD diagnoses [14]. In contrary to our approach, the described approach in [14] consisting a model based on convolutional block attention module (CBAM) [15] which is an attention mechanism based on convolution mechanism. Our approach is convolution free. They used both public and private datasets and compared their attention-based model with the same model without the attention mechanism. The model with the attention mechanism was superior detecting AMD features in most of the cases.

4 Data

We used a private dataset of the Hadassah Medical Organization's Division of Ophthalmology. The dataset consisting 40891 OCT scans taken from 616 patients. Each patient has several but different number of OCT scans taken through the years that the patient was getting treatment in the hospital (65 scans in average for each patient). Because of the fact that the dataset was built from patients, the number of healthy patients scans (control team) is very small (only 18 patients, 268 scans) in comparison to the ill patients.

4.1 Data preprocessing

Each OCT scan stored in E2E file (also called Heidelberg OCT). In order to extract images from E2E files, we used a python library called "oct converter". From each E2E file, we got different amount of image files which representing the retina in single width strip (called slice). Because of the scan quality and the format conversion, the received slice images are noisy and misaligned. We then applied some image processing techniques in order to crop irrelevant areas and to align the images to the center. Unfortunately, due to the high noise, we could not align the images with each other. Alignment between the slices could have allow us to apply temporal attention and improve our results.

Table 1: AMD stage prediction - classification method.

Method	No-AMD	Dry-AMD	Wer-AMD	Total
Binary	-	**0.85**	**0.84**	**0.84**
3-class	0.038	0.82	0.74	0.76

Table 2: Different models testing. HR and L versions were tested only with 10% parameters finetuning and their hyper-parameters were not tuned due to resources limitations.

Model	Accuracy	Recall	Precision	AUC
Timesformer	**0.84**	**0.84**	**0.94**	**0.9**
Timesformer-HR	0.79	0.85	0.85	0.8
Timesformer-L	0.78	0.86	0.84	0.79

4.2 Data labels

The data was labeled by Hadassah Medical Organization's Division of Ophthalmology stuff. The labeling preformed only for the first and the last OCT scans for each patient. This means that the majority of the data is unlabeled. The data was labeled by 12 AMD related biomarkers that can be observed by human from the OCT scan. In our work, we took advantage on the fact that some features are irreversible. If an irreversible biomarker exists in the labeled first and last scans of certain patient, we could know for sure that this bio-marker exists in all of the patient's other scans. This observation increased our labeled data from 3% (1200) manually labeled to 75% (30000) automatically labeled scans. **Note:** the current manual labels probably has a significant number of errors. Fix should be released soon from Hadassah's staff and we expect for improvement in results.

5 Experiments

For instance, we chose to focus on prediction of AMD stage. From one hand, this can raise some problem because this is a subjective feature. From the other hand we though that rescale the images into lower resolution, will make it harder for the model to observe small details. We used Timesformer smallest version that takes 8 frames in size 224x224 with pretrained models. The pretrained models are taken from the Timesformer work github and were trained on several datasets and tasks. We represent here results that predicted on the **validation** set because we want to save out test set for future work.

5.1 AMD stage prediction

The data was labeled to 0 - no AMD, 1 - dry AMD and 2 - wet AMD. Unfortunately, there were only 70 scans of no-AMD group in the training set. The performance of three class classification, even with class weighting, was poor. It reached total accuracy of 0.76 but recognized no-AMD samples only in accuracy of 0.0385. We chose to focus on binary classification for instance between dry-AMD and wet-AMD. We performed as many as 100 different versions of this experiment and we reached to total accuracy of 80%, 83% and 77% accuracy, 0.6 and 0.55 recall and 0.8 and 0.5 precision of majority and minority class respectively (Table 1).

5.2 Different models

We achieved the best results with the smallest Timesformer because we had difficulties in training the other two models Timesformer-HR and Timesformer-L (Table 2). We ran the large models under certain limitations and they seemed to have potential for producing good results. When finetune only 10% of the parameters with small batches, Timesformer-L achieved 78% accuracy and Timesformer-HR achieved 77% accuracy. However, their AUC and precision were not very high. The results achieved in single run, with no hyper parameters tuning for the large models, That means that they probably can perform a lot better.

Table 3: Models with different finetuning percentage. Percentages does not include the final MLP parameters. For example, 0% finetuning means that all of the parameters are freeze except of those of the MLP. 60% means that 60% of the last blocks parameters in addition to the MLP parameters are optimized during training and the first 40% blocks parameters are freeze. All models are initialized with pretrained model.

Model	Finetune	Accuracy	Recall	Precision	AUC
Timesformer	100%	0.8	0.79	0.92	0.88
Timesformer	**60%**	**0.84**	**0.84**	**0.94**	0.9
Timesformer	20%	0.77	0.74	0.86	**0.92**
Timesformer	0%	0.74	0.78	0.81	0.84

Table 4: Frame picking. Timesformer and Timesformer-HR gets only 8 and 16 frames respectively. We have few heuristics to choose them out of the available frames. **8-mid-frames** takes the central frames. **Down-sampling** chose a random index and takes frames in constant interval (cyclic).

Heuristic	Accuracy	Recall	Precision	AUC
Random	0.8	0.78	0.88	0.88
8-mid-frames	**0.84**	**0.84**	0.94	**0.90**
Down-sampling	0.76	0.71	**0.94**	0.88

5.3 Fine-tuning

After we initialized our model based on pretrained model, we performed finetuning of different percentages of the network. Although the pretrained network was trained on regular videos, we expected it to be a better basis than randomly initialized network. Previous works with medical data proved that using pretrained network of regular data could be useful [13]. We chose 0, 0.25, 0.4, 0.5, 0.75, 0.9 and 1 to be the optional fractions of the attention blocks that will be freeze during the training. All of the options does not include the MLP head, that remains free. The results for the model with fraction of 0.4 were the best (Table 3). Means that 40% of the attention blocks remained constant during the training and the other 60% attention blocks parameter were optimized during the training.

5.4 Frames picking

Because of the fact that our model takes only 8 frames while each scan composed from 4 - 60 frames, we had to decide how to pick the 8 frames input. We tested three heuristics: Eight-middle-frames - takes the eight frames from the center of the scan. The logic behind this heuristic is that the center of the scan is most likely covering the "fovia" which is the most critical, active and influencing area of the eye. Down-sampling - sample eight frames with constant intervals. This heuristic supposed to give frames from all over the eye uniformly. Random: get (ordered) random frames. This supposed to help us to avoid overfitting. All of the heuristic are problematic because they are missing data. Timesformer has a version that takes 96 frames, which could help us, but this model is very large and by now we could not train it effectively. The heuristic that showed the best results was the ight-middle-frames picking. It approves that the center region, that covers the fovia, indeed has a great influence. The other heuristics are not far behind (Table 4).

6 Future work

We are intend to expand our research in few directions. First, we have to get the updated labels from Hadassah. Second, we will try to perform more tests with the larger models Timesformer-HR and Timesformer-L. Third, we have to test our work on public data sets or even use it for pretrain our model. Fourth, we will consider to use Video Swin instead of Timesformer.

7 Conclusion

We presented a 3D vision transformers approach to predicting AMD features. Our model achieved adequate results and demonstrated that 3D vision transformers can be suitable for this task. The main problems that has to be solved in order to achieve competitive results are: first, better data preprocessing - the current preprocessing does not maintain the continuity of the frames of an OCT scan. Thus, the model does not use its temporal understanding abilities. Second, train model that receives input from better resolution. The current model deals with 224x224 size input what reduces its sensitivity for small details. Using Timesformer-HR or Timesformer-L probably can improve the results. Third, the data labels has to be rechecked by Hadasah's experts (They already work on an update).

References

[1] Lim, Laurence S., et al. "Age-related macular degeneration." The Lancet 379.9827 (2012): 1728-1738.

[2] Bertasius, Gedas, Heng Wang, and Lorenzo Torresani. "Is Space-Time Attention All You Need for Video Understanding?." arXiv preprint arXiv:2102.05095 (2021).

[3] Schmidt-Erfurth, Ursula, et al. "Guidelines for the management of neovascular age-related macular degeneration by the European Society of Retina Specialists (EURETINA)." British Journal of Ophthalmology 98.9 (2014): 1144-1167.

[4] Miotto, Stefania, et al. "Morphologic criteria of lesion activity in neovascular age-related macular degeneration: a consensus article." Journal of Ocular Pharmacology and Therapeutics 34.3 (2018): 298-308.

[5] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

[6] Liu, Ze, et al. "Video swin transformer." arXiv preprint arXiv:2106.13230 (2021).

[7] Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." arXiv preprint arXiv:2103.14030 (2021).

[8] Jain, Nieraj, et al. "Quantitative comparison of drusen segmented on SD-OCT versus drusen delineated on color fundus photographs." Investigative ophthalmology visual science 51.10 -(2010): 4875-4883.

[9] Chen, Qiang, et al. "Automated drusen segmentation and quantification in SD-OCT images." Medical image analysis 17.8 (2013): 1058-1072.

[10] de Sisternes, Luis, et al. "Quantitative SD-OCT imaging biomarkers as indicators of age-related macular degeneration progression." Investigative ophthalmology visual science 55.11 (2014): 7093-7103.

[11] Sun, Zhuli, et al. "An automated framework for 3D serous pigment epithelium detachment segmentation in SD-OCT images." Scientific reports 6.1 (2016): 1-10.

[12] Wintergerst, Maximilian WM, et al. "Algorithms for the automated analysis of age-related macular degeneration biomarkers on optical coherence tomography: a systematic review." Translational vision science technology 6.4 (2017): 10-10.

[13] Saha, Sajib, et al. "Automated detection and classification of early AMD biomarkers using deep learning." Scientific reports 9.1 (2019): 1-9.

[14] Yan, Yan, et al. "Attention Based Deep Learning System for Automated Diagnoses of Age-Related Macular Degeneration in Optical Coherence Tomography Images." Medical Physics (2021). [15] Woo S, Park J, Lee JY Kweon IS. CBAM: convolutional block attention module. COMPUTER VISION-ECCV 2018, PT VII. 2018;11211:3–19.