ECE 8803 Final Project

Max Beaulieu, Daniel Alvarado, and John Mac Hale
Electrical and Computer Engineering
Georgia Institute of Technology

I. Introduction

In this project, our group was tasked with predicting the presence or absence of 6 biomarkers based on an Optical Coherence Tomography (OCT) image scan of the eye, or in simpler terms, a cross-sectional image of the eye, as well as associated clinical information. The dataset provided by the Georgia Tech OLIVES lab was utilized to both train and test the model presented in this paper. This data set provided more than 78,000 training samples and 3,800 testing samples [1]. This project was initially introduced as a competition for IEEE Signal Processing Society (SPS), but has been converted into a final project for ECE 8803.

II. Problem Setup

The project revolves around the premise of discovering and learning to identify 6 key biomarkers from OCT scans. The scans are represented in the form of 504 x 496 pixel images. While some of the images are represented in the RGB color space, others are represented in grayscaled formats. For our project, all images were converted to grayscale form as no real information was gathered from the RGB format. An example of an image from the data is presented below.

Fig. 1. Sample 1 of the OLIVES OCT biomarkers dataset

For each image, the Best-Corrected Visual Acuity (BCVA) and Central Subfield Thickness (CST) were also measured and are present in the data set.

The training data contained target data in the form of a 1 or 0 on each of 6 different biomarkers, with 1 meaning the biomarker was present and 0 representing its absence. This target data was used as the loss of the dataset. Of the dataset images, only 17,591 samples had fully labeled data. As such, all other samples were ignored. The unlabeled data initially led to a lot of code issues, drastically increasing the loss as PyTorch treated it as incredibly negative numbers.

In general terms, the goal of this project was to convert the input of an image from the dataset into a six-output array that represents the presence or absence of the given biomarkers.

To achieve this goal, we initially tested the creation of our own model utilizing the principles we've learned in class to create a CNN that converts the image's pixels into 6 final outputs. However, noticing that the performance was not at the level that we wanted, we dove into modern computer vision feature recognition models and analyzed why some performed better than others. We found that these research and pre-trained models provided far stronger performance.

Notably, we evaluated the performance of MaxViT for its ability to blend modern works of transformers into the sphere of vision feature recognizers [2]. Other contestants to this competition had found success with the model and we wanted to understand why it was so powerful for this recognition task [3].

We also evaluated the performance of Meta's DINOv2 [4]. This model is made generically for all purpose visual feature identification and uses an extreme amount of data for pre-training. We hoped to explore whether leading-edge all-purpose models still performed well in very specific medical applications.

III. Data Pre-Processing

We investigated how useful processing the input images before training the model on them would impact the model's ability to identify the biomarkers present in the images. The reasoning behind this is that if the model overfits on the training data, it would not be able to generalize well for images outside of the training set, and decrease its F1 score on the testing dataset.

We first reviewed the transformations the Synapse team used on the training data [3]. We applied a color jitter transformation with brightness = 0.5, contrast = 0.5, saturation = 0.5, hue = 0.2, and p = 0.8. The goal of this transformation was to randomly change the brightness, contrast, saturation, and hue of the image to simulate the images being taken in different lighting conditions so that the model didn't focus only on any one of those parameters, but instead generalized to distinct patterns in the image that represented the biomarkers. The random resized crop transformation had a scale of (0.7, 1), so up to 30% of the image was cropped, and the image was resized to the original dimensions of 504 x 496 pixels. This transformation would allow the model to focus on a subset of

an image to be able to identify patterns easier since there is less unique data after cropping it. The images were randomly flipped horizontally with p = 0.5 so that the model could be agnostic to the horizontal direction of any patterns in the image. The last transformation was normalizing the data with a mean of 0.1706 and standard deviation of 0.2112.

When the transformations were applied to training the custom CNN model, we found that the F1 score did not improve, and in fact was slightly worse than without any transformations. A possible explanation for this is that the transformations made it more difficult for the model to converge during training in the same number of epochs, so it's possible that more than the 20 epochs we used for training the CNN model was required to see any improvements from using the transformations.

After that, we tried to investigate the transformations applied by the runner-up team in the competition, Neurons [6]. They applied a random crop, gaussian blur, random horizontal flipping, randomly changing the image's perspective (essentially moving a 2D image in a 3D space and projecting it back into a 2D image), randomly applying an affine transform (rotation, translation, and cropping), and performing basic image segmentation to detect the image's background and setting it to black (pixel value of 0) to prevent the model from training on the background noise present in the image. In addition, the authors applied a simple adjustment where they decreased the brightness of the image and increased the contrast as a strategy for denoising the image.

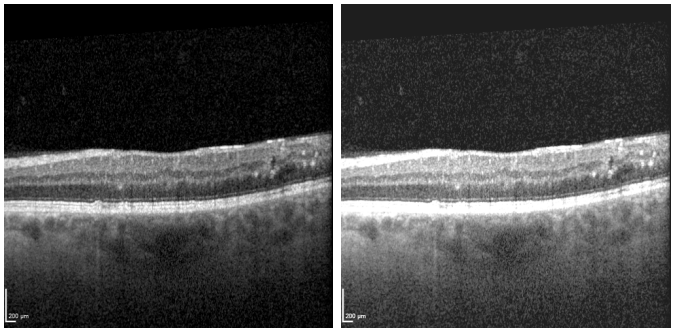

Fig. 2. Sample 50 of the OLIVES OCT biomarkers dataset before and after brightness and contrast adjustment

After training the model with the transformations used in the Neurons paper, it turned out the most effective transformation was the simple brightness and contrast adjustment. We noticed an improvement by 0.04 in the F1 score when training the CNN model with and without this single transformation which we believe is a non-negligible amount that cannot only be attributed to the shuffled order of the training data. Thus, we reused this brightness and contrast adjustment when training the other models.

IV. Model Implementations

We explored three models with the goal of comparing their performance as well as understanding why each model performed the way it did on the eye dataset. These models included a simple Convolutional Neural Network (CNN), MaxViT, and DINOv2.

A. CNN Model

Our first and simplest model was created using the knowledge we developed in this class. Knowing that an image was the centerpiece of data that we wanted to derive knowledge from, our team set upon utilizing a CNN Architecture to place emphasis on finding features from the images being presented.

Our model was rather simple, with three main convolutional blocks that took the initial 504 by 496 image and turned it into a 64 feature input into our classification (fully-connected) layer. The convolutional layers had 16, 32, and 64 filters respectively, with a 3x3 kernel and padding=1. Each convolutional block also had a max pool with a 2x2 kernel and a stride of 2. As discussed in class, the goal of each convolutional layer is to bring about more and more complexity as the data flows through the model with earlier convolutional blocks dealing with edges and corners while later convolutional blocks are detecting the more abstract features. The goal of the pooling in-between layers is to reduce the computational cost as we get deeper and deeper into the model. We don't want the fully-connected layer to be too big or the training could take a lot longer than we want. The linear layer translated the 64 feature input from the last convolutional layer into 6 outputs to represent the biomarker guess, including using a sigmoid activation function to yield probabilities.

During training, we noticed that if the training batch size was too large (we used a batch size of 200), the performance of the model suffered compared to a batch size of 32. The literature agrees with this result, the main reason being that a larger batch size causes the model to take too long to converge and a batch size that is too small results in the model bouncing back and forth too erratically, preventing convergence from occurring [7]. In addition, the complexity of the dataset, due to it being medically focused, affects the optimal training batch size.

B. MaxViT

MaxViT is a simple, fully hierarchical vision transformer backbone. As can be seen in Figure 2, it begins with a two-layer convolutional stem, where two successive 3x3 convolutions reduce the input resolution from 224x224 to 112x112. The network body consists of four stages, each of which halves the spatial resolution and doubles the channel width where, within each stage, identical MaxViT Blocks are repeated. Rather than relying on a special classification token, MaxViT applies global average pooling over the final 7x7 feature map, followed by a linear classifier, which streamlines the architecture and leverages global spatial information efficiently.

Each MaxViT Block mixes convolution and attention in three steps.

1) **An MBConv module which comprises of an expansion Conv 1x1, depthwise Conv 3x3, squeeze-and-excitation, and projection serve to enrich local**

feature extraction and indirectly encode positional information. [2]

2) **A block self-attention layer partitions the feature map into non-overlapping PxP window where P = 7, and applies relative attention within each window.**
3) **A grid self-attention layer reorganizes the feature into a GxG uniform grid, where G = 7, to perform sparse, dilated attention across grid cells. Both attention layers run in linear time and use a small two layer MLP afterward.**

This design allows each MaxViT block to see globally throughout the network, yet retain the fine grained locality essential for detecting small, texture-based patterns.

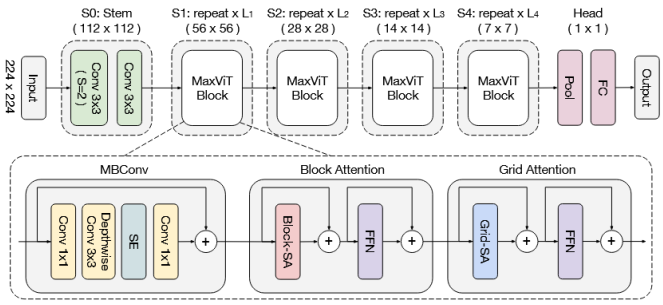

Fig. 3. Structure of the MaxViT hierarchical architecture

Because standard full self-attention scales quadratically with input size, many pure-Transformer backbones struggle with high-resolution medical scans. MaxViT overcomes this with its linear-complexity attention which makes it feasible to ingest larger OCT scans without losing clinically relevant detail. In the IEEE VIP Cup 2023 solution for OCT biomarker detection [3], the authors leveraged MaxViT's convolution and strided-attention layers to excel at local biomarker identification (e.g., Intraretinal Fluid), while pairing it with a pure-attention model (EVA-02) that better captures global biomarkers. Their success underscores MaxViT's strength in mixed local-global pattern recognition for ophthalmic image analysis.

C. DINOv2 Model

DINOv2 is the second version of Meta's DINO self-supervised learning architecture made for general-purpose image feature recognition. The model is considered a foundational model in image classification and feature recognition due to its strong performance of over 83.5% accuracy against the ImageNet dataset, as well as its easy adaptability to downstream tasks [5].

The DINOv2 model is built on the idea that learning should generally be done in a task-agnostic way, with post implementation fine-tuning on the task at hand. Therefore, the DINOv2 model used was pre-trained on 142 million images and 20 epochs of the full data was additionally used to modify and fine-tune the weights of the model.

The DINOv2 model uses a Vision Transformer (ViT) backbone. In a ViT block, the image is broken up into 14 x 14 pixel patches, processing the sequence of patches using a self-attention block that allows multiple layers to learn from each others representations. It, in essence, allows weighted mixture of how every layer is perceiving the image and aggregate information. The architecture is completely convolution-free. Instead, the model produces 2 global crops of the image as well as 10 local crops that it uses to figure out specifics of the features being targeted.

To train, a student-teacher self-distillation occurs where a teacher provides the model with expected targets with variable learning rates. The teacher model is exactly the same as that of the student except that its weights are an exponentially-moving average of the student's weights. The teacher only sees the two global crops while the student model sees all 12 crops. This structure allows feature recognition to be learned without any human-supplied labels.

The model has performed well in previous image recognition tasks due to its massive data size of 142 million images, its transformer architecture giving global attention, and its teacher-student training methodology. The multi-crop feature of the model also gives it an advantage in capturing viewpoint-agnostic behavior.

V. Performance Analysis

A. Why F1-Score?

Over the course of the semester, we've learned of various performance metrics that are utilized to evaluate a network's performance including Accuracy, Precision, and Recall to name a few. Phase 1 of the OLIVES VIP Cup project tasks us with maximizing the macro-averaged F1-score on the validation data, but we believe it to be important to highlight why this is the parameter we must pay attention to.

1) Accuracy: Accuracy will count all correctly classified samples, including the abundant true negatives. As the dataset is highly imbalanced, a naive model that predicts the majority class for all instances can achieve high accuracy while completely failing on minority classes. Thus, using accuracy as a performance metric is uninformative in this case.

2) Precision and Recall: Optimizing precision alone encourages the models to be conservative, which has the potential to miss many true positives. Conversely, optimizing recall alone can flood results with false positives.

3) ROC-AUC: ROC-AUC evaluates ranking ability across all thresholds but does not correspond to performance at the specific decision threshold used for classification submissions. Moreover, ROC-AUC can be overly optimistic in imbalanced settings because it includes true negatives in its calculation, which are in abundance in our case.

In contrast to the aforementioned performance metrics, F1-score combines the harmonic mean of precision and recall which penalizes both false positives and false negatives. Averaging per-class F1 scores ensures that models must perform well on every biomarker and not just the most frequent ones, thus aligning with the project's objective of balanced, clinically-meaningful detection.

B. Evaluation

Following training through the entire dataset, the F1-score was calculated for all 3 models, utilizing the equation below:

$$\text{F1 Score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

The results of this calculation is presented in table 1.

TABLE I
COMPARISON OF MACRO-F1 SCORES FOR DIFFERENT MODELS

Model	Macro-F1 Score
Simple CNN	0.5883
MaxViT	0.6727
DINOv2	0.5163

The simple CNN model had an F1-score that was rather low compared to the MaxViT implementation, implying a level of underfitting. We believe the model likely had too shallow of an architecture with the aggressive 2x2 pooling likely reducing the total amount of information available to the fully-connected layer by too much and too quickly. Instead of having just 3 convolutional layers, the models from papers have tens to hundreds of layers that all have specific purposes and targets.

In addition, the DINOv2 model performed worse than the custom CNN model. This demonstrates that the convolution-free architecture of the model may not lend itself well to domains such as medicine that contain biological patterns that can be learned from convolution operations due to the presence of local patterns that can be extracted through convolution operations. This is possibly why the custom CNN model which was not trained on any previous data was able to outperform the DINOv2 model which was previously trained on the Imagenet dataset as well as the biomarker training data.

The highest score that the MaxViT model was able to achieve was through the maxvit_xlarge_tf_512 model that supports input images of 512 x 512 pixels. However, it only increased the F1 score by 0.01 compared to the maxvit_tiny_tf_384 model, and the training and inference times were much longer due to the much larger model size (four times the amount of parameters). Thus, the smaller MaxViT model is preferable due to the low drop in accuracy compared to the speedup achieved.

The CNN model was quickest to train and run inference on compared to MaxViT and DINOv2, so there's a tradeoff between speed and accuracy for the CNN and MaxViT models.

VI. INSIGHTS AND ANALYSIS

We did not implement complex regularization in the CNN model, hoping to get more of a baseline that can identify performance boosts from models that have been more specialized for these tasks. Introducing dropout regularization likely would have allowed the model to notice more complex features of the data.

Additionally, the winning models for the 2023 IEEE SPS VIP Cup [3] [6] utilized a combination of top-performing models by ensembling them together to average the model outputs to achieve more accurate predictions and increase the combined overall F1 score. Reflecting on the performance of individual models, it makes sense that certain model architectures are better at detecting certain biomarkers than others. If two models agree on the presence or absence of a biomarker, then the prediction will be stronger (probability is closer to 1 for presence and closer to 0 for absence). If the two models disagree, the averaged probability is biased towards the model which made a more confident prediction (probability is further away from 0.5). Thus, the probabilities are effectively strengthened and the combined models have better accuracy which is corroborated by the papers submitted to the competition.

We also could have looked into the EVA-02 model [8] which was cited as having performed better at predicting global features which convolution alone would have a harder time detecting [3]. The EVA-02 model utilized EVA-CLIP as its teacher as part of a process known as knowledge distillation. The difference in MaxViT's performance compared to EVA-02 is attributed to the difference in attention mechanisms. The suggested reason is that MaxViT uses sparse attention to only look at a subset of pixels while EVA-02 uses true attention which means it uses all of the image pixels, so it makes sense why EVA-02 would be better at detecting features that span the entire image instead of a small local region.

Another model we could have researched is CoCa [9], the model with the highest top-1 accuracy on the Imagenet dataset [10]. CoCa uses contrastive learning to differentiate between similar and dissimilar data points. CoCa by default doesn't use convolutions so it is more likely to perform better at detecting global features rather than local features. However, there are CNN-based versions so we could include that as a parameter to adjust to see if adding convolutions to CoCa would improve detection of local biomarkers.

The GitHub repo can be found at: https://github.com/maxbhockey3/ECE8803_Final.

REFERENCES

[1] OLIVES Dataset, [Online]. Available: https://huggingface.co/datasets/gOLIVES/OLIVES_Dataset

[2] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "MaxViT: Multi-Axis Vision Transformer," *arXiv preprint arXiv:2204.01697*, 2022. [Online]. Available: https://arxiv.org/abs/2204.01697

[3] H. A. Z. S. Shahgir, K. S. Sayeed, T. A. Zaman, M. A. Haider, S. S. R. Jony, and M. S. Rahman, "Ophthalmic Biomarker Detection Using Ensembled Vision Transformers," *IEEE SPS VIP Cup*, Team Synapse, 2023. [Online]. Available: https://bpb-us-e1.wpmucdn.com/sites.gatech.edu/dist/4/3061/files/2023/10/Synapse.pdf

[4] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, *et al.*, "DINOv2: Learning Robust Visual Features without Supervision," *arXiv preprint arXiv:2304.07193*, 2023. [Online]. Available: https://arxiv.org/abs/2304.07193

[5] D. Pasanisi, E. Rota, A. Zaccaria, and S. Ierace, "Transformed-based foundational models in Computer Vision: an industrial use case," *Procedia Computer Science*, vol. 232, no. 2, pp. 823–830, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050924000826

[6] M. Islam, M. Abtahi, M. Chowdhury, M. Hasan, A. Quadir, and L. Aktar, "Ophthalmic Biomarker Detection with Parallel Prediction of Transformer and Convolutional Architecture." Accessed: Apr. 24, 2025. [Online]. Available: https://bpb-us-e1.wpmucdn.com/sites.gatech.edu/dist/4/3061/files/2023/10/Neurons.pdf

[7] I. Kandel and M. Castelli, "The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset," ICT Express, vol. 6, no. 4, May 2020, doi: https://doi.org/10.1016/j.icte.2020.04.010.

[8] Y. Fang et al., "Eva-02: A visual representation for neon genesis," arXiv.org, https://arxiv.org/abs/2303.11331 (accessed Apr. 24, 2025).

[9] J. Yu et al., "Coca: Contrastive Captioners are image-text foundation models," [2205.01917] CoCa: Contrastive Captioners are Image-Text Foundation Models, http://export.arxiv.org/abs/2205.01917 (accessed Apr. 24, 2025).

[10] "Papers with Code - ImageNet Benchmark (Image Classification)," paperswithcode.com. https://paperswithcode.com/sota/image-classification-on-imagenet